International Labour Organization | CINTERFOR

ILO/Cinterfor Notes

Nº 11 February 2022

Date: 2/17/2022

Three insights derived from three studies of big data in the labour market.

This note quickly analyzes the evolution of data management on employment intermediation and reviews the characteristics of three national studies in which, in order to establish skill gaps, data recorded in the databases of public or private labour intermediation companies have been extracted. Emphasis is placed on the evolution of the data; from the typical form completed by hand (1970s), which then moved to a computerized application (1990s) and which, at present, is in the world of the Internet or "the cloud", as they say. The central idea is to show the enormous potential that data in the cloud has for analyzing supply-demand mismatches in the labour market - from labour intermediation - as well as the limitations still present, based mainly on three studies developed by ILO/Cinterfor in 2020 and 2021 in Uruguay, Dominican Republic and Paraguay.

Employment intermediation: from a linear supply-demand encounter to a multi-relationship of data and variables.

A job search always has two sides. The labour demand, originating from the business activities that require skills and competencies available in people, and the labour supply, precisely the people who put their skills on the market to get a job.

Job demand is expressed in information, variables and data. Usually the title of the position gives the most notorious information, "required... wanted... important company needs..." are the typical headings of a vacancy published in a job advertisement. In addition to the name of the position, other information is added, among which are usually, a brief description of the job, its functions, the place where the company is located, educational level required and some other requirements.

In the early years of public employment services, all this information was written down by hand, on long forms. One of the most important jobs in the employment offices was to interview the job seeker and fill out the form. In the same way, and usually by telephone, the vacancy data was received from the job seeking companies.

The opportunities to systematize this information were few, the data provided by the periodic monitoring and results reports referred to the vacancies received and the registered providers, the placement rate achieved and some other data classified by age, sex, economic sector, educational level and salaries offered and/or demanded.

With the advent of the Internet, the first thing that was automated was the match or meeting between supply and demand. This work was initially done by hand, comparing the files of the offerers against the vacancies that were received and almost always guided by the good judgment of the person acting as intermediary. The use of occupational classifications[1] was unavoidable as the only way to find the ideal match for a certain vacancy. This meant that every vacancy or every job search was given an occupational code and sometimes two or three optional codes to facilitate this "encounter" or supply-demand match.

The automation of the match process accelerated rapidly in the 1990s and, subsequently, the production of reports and the consolidation of national data were also automated. However, in several countries, there is still no national database or full aggregation of job supply-demand data.

[1] Normally the national classification of occupations in use in each country, strongly inspired by the ILO's International Standard Classification of Occupations (ISCO).

The interconnection between the data collected by the employment offices in a large national database is an objective for several countries and then, the exploitation of all these data for their analysis in terms of skills gap and anticipation of demands.

The advance in the use of online vacancy databases has opened a door today on the possible advent of a new tool for analyzing the labour market.

These studies known as big data analysis present certain advantages such as their timeliness and immediate availability of information, their low cost and the enormous capacity to process data and find trends; however, the massive data analysis tools still face enormous challenges, mainly due to the representativeness of the data (biased by the type of users that preferentially use public and private labour intermediation services), the "purity" of the data due to the innumerable range of names, names, definitions and also the lack of identification of a vacancy.

Nowadays, massive supply-demand data analysis is well known in companies such as Burning Glass Technologies, JANZZ.Technology, LinkedIn, LinkedIn and institutions such as CEDEFOP and various ministries of labour in ministries of labour in Europe. The ILO has also initiated the analysis of big data to measure imbalances.

However, in Latin America there are still few web scraping applications or massive analysis of supply-demand data in the labour market.

Supply and demand data: inputs for the anticipation of training needs

The online information stored in the databases on supply (job seekers) and demand (open vacancies in companies) in a temporal analysis perspective, opens a wide range of possibilities to study variations in demand and to scrutinize possible trends on growth, decrease or stagnation.

By analyzing the evolution of a certain group of vacancies over a period of two, three or more years, you will have a trend overview that can be illustrative. Also the irruption at a certain time of demand for new vacancies or some of little movement in a certain market, can be a clue to a growing occupational profile.

Clearly, the anticipation of demand and the measurement of skill gaps can be supported by massive data analysis mechanisms in public and private labour intermediation databases.

However, there are still limitations to the full use of this approach, such as the representativeness of the vacancies channeled through the databases, the structure included in their publication, their relationship with recruitment practices and their possible bias due to the size of the informal economy. With these ideas in mind and to broaden the knowledge and applicability of new tools, **ILO/Cinterfor** has been supporting the use of big data analysis to better understand the supply-demand interaction and to facilitate evidence-based decision making. In the following, we will briefly describe the studies conducted and propose four central ideas derived from them.

Three big data studies on labour intermediation

These studies were carried out in Uruguay (2019), Dominican Republic (2021) and Paraguay (2021). In all cases, these were ILO and **ILO/Cinterfor** technical assistance processes for the Ministries of labour. The different consultancies that carried out the studies had the support and technical guidance of **ILO/Cinterfor**; therefore, the methodological approaches to the analysis were deliberately different, which enriches the knowledge base and allows for a more comprehensive analysis.

Main characteristics of the Uruguay study (January to September 2020):

In Uruguay, within the framework of a project with the Ministry of labour and Social Security, **ILO/Cinterfor** stimulated the implementation of a big data analysis exercise. It was carried out with a database of an online job portal: BuscoJobs[2], which analyzed labour demand and compared it with indicators of supply availability to arrive at the main mismatches. An important feature of this database is its consistency over time in terms, for example, of correlating positively with the activity rate, the employment rate of the economy and with indicators such as the total number of jobs with effective Social Security contributions.

This study also had the advantage of making it possible to link labour demand and supply data from a single database with high representativeness in terms of published vacancies. In some cases of other experiences, scraping only covered labour demand, as in the Paraguay study.

Among the most relevant features extracted from the study are:

[2] For reasons of availability and representativeness of information, the analysis considered only data from this database, although it did facilitate some comparisons with other online portals.

- The largest number of jobs in demand tends to be concentrated in urban areas.
- Most companies do not publish age (77%) and/or gender (87%), including salary (79%), as relevant dimensions of the vacancy.
- The average age of job applicants is around 30 years old (within the group: 45% men, 55% women).
- The vast majority of vacancies (87%) are positions with no staff or supervisory duties.
- 78% of the vacancies require full-time work.
- 67% of vacancies originate in medium-sized or large companies.
- 17% of vacancies are generated by "employment agencies" that usually offer outsourced placement services.
- Most of the labour demand is in the commerce and services sector (29%), followed by professional activities (18%) and information and communication (12%).
- Only 2% of job seekers were in the information and communication sector, which generated 12% of job vacancies.
- Another imbalance is presented in professional, scientific and technical activities, which only represented 5% of the supply compared to 18% of the total vacancies generated by this sector.
- By occupations: demand was concentrated in services and salespeople.
- If we group the occupations related to information technology, science, engineering at the professional and technical level, we will have 16% of the demand, only surpassed by a single occupational group in demand -without requiring a high level of education- that of 24% of the salespersons.
- By educational level, there is a marked gap in the young population seeking jobs with low educational level, less than 20% of the vacancies that requested educational level accepted incomplete high school while, on average, in the capital of the country, 38% of the unemployed are looking for jobs without finishing high school and 45% of the employed who are looking for jobs and have not completed this educational level either.
- In terms of languages, one in four vacancies required some level of English proficiency.
- By specific skills required in the vacancy, it was found that 41% asked for marketing, communication and commerce and 30% for information technology (easily associated with digital skills).
- The second half of 2020 showed declines in demand in several sectors (accommodation, services, education, arts and recreation) due to the COVID-19 crisis.
- The study generated an indicator for the degree of fit between applicants and vacancies, with 100 being the perfect degree of fit, reaching a national average of 49 points. The greatest mismatches were in areas outside the capital, or in the case of requiring post-secondary education levels and when other languages were required. By sector, the largest mismatches were in information and communication, professional activities and teaching, health and social services, and health and social services.
- An interesting percentage weighting was that ICT occupations accounted for 12% of the vacancies, but only 6.7% of the training course offerings surveyed in the study.

Highlights of the big data study in Paraguay

In Paraguay, in the framework of the support that the ILO Southern Cone office and **ILO/Cinterfor** offer to the Ministry of Labour, a web scraping analysis was developed in 2020 and a second edition in 2021[3].

The study was based on a methodological proposal that dimensions the skills gap based on mismatches perceived by the behavior of occupations in the supply-demand interaction.

Vacancies are identified in occupations with high turnover, hard-to-fill vacancies and neutral or normal behavior vacancies (they occur with low frequency and are filled without delays above the averages). As an interesting product, the study added a display board for a better representation of the data.

The gap is configured centrally in hard-to-supply occupations and a "frictional unemployment" [4] effect is also attributed to high-turnover occupations. For occupations that mix features of difficulty of provision and high turnover, the concept of critical occupation is configured.

[3] The results refer to 2020 as the aggregate information for 2021 was minimal and did not change the trends. The studies are published at: https://www.oitcinterfor.org/brecha_BigData

[4] Friction is characterized by the time elapsed between the generation of the vacancy and its filling; an effective information system tends to reduce frictional unemployment.

As an information base, public and private employment portals were accessed through scraping and the available information on job vacancies was downloaded. This information was complemented with data from household surveys, especially to characterize the labour supply, whose data were not accessible in the scraping process[5].

- For the scraping exercise, portals such as: Paraempleo, Redtrabaje, Opción Empleo, BuscoJobs, CompuTrabajo, LinkedIn, Pivot and Redi were reviewed in the period from June to November 2020 (strongly marked by COVID-19) for 3185 vacancies. We were able to purge information on some 188 occupations of which 54 were critical.
- The methodology for defining critical occupations has been used in several countries with World Bank studies, including Indonesia, Malaysia, Australia, Ireland, the United States and Colombia, among others.
- The central idea is to publicize the gap so that decision-makers and skills and competency development and training programs can focus, in the short term, on closing the gap in such critical occupations.
- Exercises to define critical occupations are useful for re-evaluating, redesigning or developing new vocational training programs, encouraging apprenticeship and in-company training programs, and stimulating migration admission processes to critical occupations.
- It is based on the definition and calculation of a set of indicators to support the characterization of occupations; these are based on different information such as frequency of publication, incentives offered, repetition of vacancy notices, participation of the vacancy in the set of demands, among others.
- This work provided a first methodological guide that is available for consultation on the ILO/Cinterfor website.
- Among the occupations with the highest turnover are ISCO-08 in the group of services and salespersons in stores and markets; followed by the group of journeymen, operatives and craftsmen in mechanical and other trades and then by the group of technicians and mid-level professionals.
- At the high skill level, the most in-demand occupations are *Android* programming and development, support analyst, systems developer, IT coordinator (32% of vacancies) and human resources analyst.
- At the medium qualification level, occupations in the digital area such as DevOps Architect, Java Architect, Python Architect, Ruby Tech Lead and Scrum Master predominate (25% of vacancies).
- At the low qualification level are sales consultants and promoters (22%) and real estate sales consultants (18%).
- A brief comparison of the structure of the demand with the structure of the educational and training offer shows that the percentage weight of graduates in ICT professional occupations is 1.47% for ICT technicians is 2.26%, while the weight of vacancies in this area for professionals is 21% and for technicians 2.66%. There seems to be a preference in ICT vacancies for graduates of the professional level.
- A dynamic information panel was produced from this exercise and is available on the ILO/Cinterfor website.

Findings and features of the study in the Dominican Republic

In the Dominican Republic, within the framework of **ILO/ Cinterfor** technical assistance to the Ministry of Labour, a web scraping study was conducted and completed in early 2021. The study available on the ILO/Cinterfor website, included an analysis of the labour market insertion of graduates of the National Institute for Technical and Vocational Training (INFOTEP).

The period from 2018 to the first quarter of 2021 was covered.

The study also looked into the social security database of the Labour Registry System (SIRLA) to try to find some traceability between INFOTEP course graduates and their subsequent employment. This source of information can still be better exploited, with more detailed access to available information and the opportunity to improve the massive analysis of data, the report emphasizes.

11 thousand vacancies were compiled from the databases of "EmpléateYa" with the highest number of vacancies analyzed, the official employment portal of the Ministry of Labour, and the portals "Tu Empleo RD" and "CompuTrabajo". The period from 2018 to the first quarter of 2021 was covered.

The study also looked into the social security database "SIRLA" to try to find some traceability between INFOTEP course graduates and their subsequent employment. This source of information can still be better exploited, with more detailed access to available information and the opportunity to improve mass data analysis, the report highlights.

[5] On general, offer data are protected as personal data and are not available for access, except when expressly provided by the database concerned. .

- The downloading of unstructured data6 required a major work of debugging and cleaning of information, followed by classification and categorization using ISCO 08 to conclude in indicators on labour market dynamics.
- The debugging, homologation and assignment of ISCO 08 codes was supported by the statistical software "R" and "Stata". It is noted that when accessing different portals, duplicate vacancies were "cleaned", i.e. the same vacancy appearing in more than one portal.
- There is an inversely proportional relationship between the trend of the unemployed population (increasing) and the enrollment of applicants in the employment service (decreasing). No clear causes were found to explain this.
- The report gives a very good description and adds an annex on the scraping, cleaning, debugging and coding techniques used.
- There is a certain gender preference in job searches. While in the administrative, pharmaceutical and health sectors the majority are women, the opposite occurs in the energy, automotive and construction sectors.
- The highest number of vacancies in the period analyzed were in the occupational groups of professionals, scientists and intellectuals, followed by mid-level professional technicians and service workers and salespersons.
- Nearly half of the vacancies require a high level of qualification (45.5 %), 34.25 % medium and 18.8 % low.
- In this case it was found that up to 60% of vacancies require up to two years of experience. If you go down to one year, 30% of the vacancies require it as minimum experience.
- The most common words or phrases in job descriptions are "teamwork", "relationship building", "leadership", "people management", "organization" and "efficient communication".
- Among the digital skills the most important are the knowledge of big data, and with a much smaller participation the knowledge of programming, data analytics and digital platforms.
- Most of the vacancies do not specify a particular profession and refer to "professional degree" for post-secondary education. The most commonly mentioned areas are "advertising and marketing", "systems", "administration" and "business management".
- In the analysis of the employability of INFOTEP graduates based on SIRLA, it was found that the data are very restricted for follow-up, especially because this database mainly contains registered or formal employment. This resulted in a low level of placement of graduates in this database.
- However, a clear trend was that 71 % of the graduates registered in SIRLA came from Dual Training, 60 % from technical teacher training.

Three thoughts on the three studies

1. Towards the massive use of big data for gap measurement and demand anticipation. A long way to go.

It is clear that the data used suffer from deficiencies in terms of structure, problems of duplication and representativeness; therefore, a cleaning process is needed, as well as verification of their quality and applicability.

Some issues about the use of private data and how to preserve privacy during the use of these techniques are still unclear. In most cases, this is solved by the basic rule of not allowing access.

In the cases observed, and most probably in the rest of the region, the scope and coverage of the vacancies published in the intermediation systems may be affected by various biases, such as urban concentration, informality, the language used and redundancy in the publication as a resource to attract supply, among others.

Nevertheless, the use of supply and demand information from intermediation databases has great potential as additional information for a better understanding of the functioning of the labour market, and to complement valuable information obtained from other sources such as surveys.

It is necessary to continue monitoring the continuous improvement of artificial intelligence tools to obtain a better use of the available data. The great challenge is to convert data on job offers and demands into knowledge applicable to policy design and evidence-based evaluation.

[6] For the scraping of the accessed portals, algorithms were built using Python and its Selenium and BeautifulSoup libraries.

2. The importance of continuing to improve methodologies and tools.

Still the information on vacancies obtained from the intermediation portals is very unstructured, it may respond to the minimum requirements of the portal or it may simply define a few clues about what the employer is looking for. The variation between the types of vacancy announcements is enormous, the use of terms to refer to the skills required is often general and ambiguous, in many occasions other key parameters such as salary offered, level of education are not defined, in others restrictive features such as the area of residence or the length of the working day are added [7].

Also, the greater or lesser number of intermediation portals implies a certain segmentation of the market in terms of the type and level of vacancies that can be found and according to what employers assume to be the success of posting a vacancy on this or that site. The databases have different structures and even, depending on the country in question, one can find the prevalence of some intermediation system in terms of capturing vacancies from the market.

In this line, for example, there is still no uniformity of criteria on whether the information on online vacancies could or should be centralized in an analysis area of the labour ministries to facilitate collection, standardization and analysis. There are no clear rules on access to intermediation data and, in many cases, the data are totally or partially blocked for automated collection.

For all these reasons, there are legal and methodological challenges to make the massive analysis of supply and demand data a better tool that can complement other existing tools. It is necessary to continue learning about the possibilities and limitations of such a methodology in the region and to support the countries in developing the capacity to analyze their massive data on labour supply and demand.

Also along these lines, the idea of seeking public-private partnerships is welcome. Large private databases can be a good source to feed public policy decisions. In many cases they have a regional vision and can aggregate indicators of mobility, transferability between regions and skill levels, changes in occupations by country within a region, among others [8].

3. Turning big data into labour market information. New horizons in the use of online data for gap measurement and demand anticipation [9].

The first efforts to work with vacancies extracted from online advertisements in Europe started at the European Center for the Development of Vocational Training -CEDEFOP-, in 2014, in a context such as the European one, with a greater structuring of information. Six years later CEDEFOP has set a series of research objectives in areas such as:

- Emerging trends in competencies and skills that can be the basis for structuring hierarchies or taxonomies.
- Transferability of required skills between occupations and sectors.
- Demand for new digital skills and their interaction with occupations.
- Indicators on mismatches and gaps in skills and occupations.
- Demand for education and job training and new skills.

Along the same lines, the European Training Foundation -ETF- is advancing in a project that favors the conversion of big data into information on the labour market. After implementing several pilots in selected countries (starting in 2019), progress is being made in defining cooperation mechanisms with statistical offices and other research actors to generate a shared methodology that can then be transferred through capacity building actions in technical teams.

Research in Italy on occupations in the ICT sector is looking into the correlation between "hard" (technical) and "digital" skills and the degree to which a job may be replaced by automation. It also includes the complementation or enrichment of skills descriptions and taxonomies using artificial intelligence techniques to analyze the reference to occupations in intermediation platforms.

[7] In the chapter: "From big data to Smart data: "The misconception that big data provides useful predictions about skills" of the document "The feasibility of using big data in anticipating and matching skills needs", the experience of JANZZ.technology, a large company dedicated to this analysis, is presented (Pages 24 to 33).

[8] One example is the LinkedIn platform: Economic Graph. Or the studies offered by Burning Glass Technologies. Or those of JANZZ Technolog

[9] The trends referred to can be consulted in depth in: The feasibility of using big data in anticipating and matching skills needs. ILO. 2020.

This research aims to develop a graphical database to display information on the European labour market based on online vacancy information, complementing taxonomies such as ESCO, but with real-time information.

But also, the enormous statistical imperfection of the available data (in terms of quality and consistency, for example) in the labour intermediation databases is one of the main obstacles to overcome in order to ensure that database scraping methods and the extraction of information on job vacancies and job offers can be a reliable mechanism for predicting demand and estimating gaps.

In short, the existence of static repositories of information on the labour market, such as taxonomies and the definitions and contents of occupational profiles, will be progressively complemented, impacted and, perhaps, replaced by classification processes based on online information, refined, organized and presented with the support of artificial intelligence algorithms, which can be published and made visible to end users.

Fernando Vargas

Senior Specialist ILO/Cinterfor

Contact

ILO/Cinterfor
Uruguay Avenue 1238
Montevideo, Uruguay
P. C. 11,100

+598 2902 0557
oitcinterfor@ilo.org