Workshop on Big data and the future of work

Meeting report

Workshop of the joint EU-ILO Project “Building Partnerships on the Future of Work”
Workshop on big data and the future of work

Meeting report

February 2021
• The Joint Research Centre of the European Commission, the Labour Market Information Council of Canada and the ILO organized a research webinar, entitled “Big data and the future of work”. This was organized as part of the EU funded project “Building partnerships on the future of work”.
• The webinar took place on Thursday, the 3rd of February 2022 via Zoom.
• Experts from different institutions came together to discuss the possibilities of making use of big data sources to analyze ongoing future of work trends in both developed and developing economies.
• Research based on different types of big data such as online job vacancies, geospatial data sources and others was presented, and their use for policy purposes was discussed.
Session 1: Online job vacancies and CVs

Introductory remarks

Introduction by Guillaume Delautre (Employment Policy Department, ILO)

Mr. Delautre kicked off the proceedings by welcoming the attendees and participants in the workshop. He noted that “Building partnerships on the future of work” is a project funded by the European Commission (EC), and implemented jointly by the ILO and the Joint Research Commission (JRC) of the EC. The aim of this project is twofold: to undertake research in the domain of the future of work and to build bridges between EU and non-EU countries on the topic of the future of work.

Mr. Delautre indicated that the aim of the webinar was to bring together experts from various institutions to get their viewpoints on the usage of big data source to analyze future of work trends around the world. Although labour market specialists have relied heavily on traditional sources, such as Labour Force Surveys (LFS), such sources may not paint a complete picture of changes. Alternative sources can help in such understanding. In the present workshop, recent research papers as well as new emerging data sources were to be discussed. Mr. Delautre then mentioned that the workshop would be divided in two sessions, and yielded the floor to Ms. Olga Strietska-Iлина from the ILO Employment Policy Department.

Opening remarks by Olga Strietska-Iлина (Employment Policy Department, ILO)

Ms. Strietska-Iлина welcomed participants and attendees on behalf of the ILO, the JRC and the Labour Market Information Council (LMIC) of Canada. Ms. Strietska-Iлина referred to a previous workshop organized by the ILO in 2019 on the feasibility of the use of big data for skill needs anticipation and matching. Ms. Strietska-Iлина then indicated that the topic of the current workshop was in fact the measurement of the supply and demand of skills, tasks, as well as jobs by utilizing online job vacancies data (with the help of big data analytics). Big data offer several advantages, including the fact that they are readily available, and are not time and labour intensive, such as the collection of LFS. Moreover, such data are by definition in real time, and offer a high level of granularity and detail. Some limitations include coverage and representativeness, which are felt in particular either in the case of low-skilled jobs, or in developing countries. Ms. Strietska-Iлина then laid out the agenda of the workshop for the first session, which would consist of 4 presentations.

Presentations

Tony Bonen (Labour Market Information Council of Canada)

Mr. Bonen’s presentation, entitled “Challenges and opportunities in leveraging online job ads to track labour shortages” focused on vacancies, job postings and their relation to shortages. There can be a potential trade-off between real time data which may have sample selection issues, and official statistics such as surveys that are slower and can have sample size limitations, but are more representative. Mr. Bonen started the presentation by noting that labour shortages have become a major policy concern in Canada. There is, however, not a reliable way to track such shortages. Several recent signs are positive- there has been an expansion in the demand-side, as employment seems to have bounced back, particularly among core aged workers, and unemployment and labour force participation rates have returned back to pre-Covid levels. From the labour supply side, growth in the working age population does not seem to have grown in the same way, as immigration levels have not recovered since before Covid. Mr. Bonen then transitioned to explaining the concept of labour shortage, which is characterized by a lack of viable candidates for specific job openings, at a specific
time and place. Such shortages are hard to measure as there is generally a paucity of data on unmet demand, as well as available supply at a granular level. Latest data shows a recent hike in labour market tightness (the vacancy/unemployment ratio) at a macro level, although data from 2020-Q3 was not collected as the ability to collected data from firms was limited at the time. Moreover, the data is quarterly in nature, and is available only at a high macroeconomic level. Policy interventions are then unable to be targeted to specific sectors or skills that may be driving such tightness. The two official data sources for unmet demand and labour supply are the Job Vacancy and Wage Survey (JVWS) and the LFS respectively. An advantage offered by the JVWS is that it contains data on vacancy duration, but is only available at quarterly frequencies as well as with a 2-3 months delay. Data from the LFS is available quicker (only 1 week lag from reference month), but the sample size among unemployed is small.

One measure of excess labour demand which would be both timely and granular was the balance index, which is defined as the ratio of the growth of job postings to the growth of potentially available labour supply over 3 months. This index is measured at the occupation level. 1 Given that occupations differ in baseline/systematic levels of job turnovers, considering this measure in ratios of growth rates accounts for these differences by focusing solely on changes from baseline levels. This measure suggests that in 2020, job postings fell rapidly, while the potential labour supply increased. The opposite phenomenon took place in 2021, where job postings increased precipitously, but labour supply did not. A more granular analysis suggests that occupational groups requiring college or apprenticeships as well as high school or less were driving this increase in growth in vacancies relative to labour supply. Occupations that require higher education levels are not characterized by this phenomenon to the same extent. Validations of this new measure with existing data from the JVWS include checking for correlations in wage growth, among others. Further refinements include applying the index to sub-national and sub-provincial regions, and non-linearity considerations in the index among others.

Matteo Sostero (Human Capital and Employment Unit, JRC)

Mr. Sostero’s presentation entitled “The Professional Lens” focused on how online job advertisements could be informative about the composition of jobs. At the outset of the presentation, Mr. Sostero mentioned that the role of the EU Skills Intelligence was to analyze data on the nature of work as well as the demand for skills, gleaned from big data in the form of online job advertisements (OJA). However, is OJA data reliable for skills intelligence, and is it representative of all job types? The research carried out by the JRC included comparing data from OJA (Burning Glass Technology) with structured occupational task data [JRC-Eurofound Task Database] sought to answer this question. The main findings were that i) the skills vocabulary in OJA was both rich and growing, expressed in the language used by employers and that ii) there was a bias in OJA favouring professional occupations and tasks. Consequently, OJA could serve as a useful complement, but not as a substitute for conventional surveys of tasks and skills.

OJA suffer from several limitations, including a lack of representativeness, and explicitness in ad text, among others. Data from Burning Glass Technology suggest that there is a wide and persistent gap in number of skills mentioned by ad, across major groups (SOC-1), with high-skilled occupation groups typified by more skills mentioned by ad. One of the questions raised was whether this difference reflects a true gap in skills requirement, or simply a measurement bias in skills-counting? Regarding the over or under/representation of occupations, the work from the JRC compared the occupation volume from the UK Labour Force Survey (LFS) with the number of online job ads by occupation at the granular level of SOC-4. The authors find that professional occupations seem to be overrepresented, whereas lower skilled occupations are underrepresented. After, developing a conceptual framework to understand and measure workplace skills in

1 More precisely, not only a particular occupation, but also similar occupations are considered in this measure, to avoid low sample size problems. The grouping is based on broad occupation groupings.
terms of tasks, Mr. Sostero and co-authors built occupation task profiles, and correlated indices of task content between the OJA data and the JRC-Eurofound Task database. They found, on average, that task intensity measured in OJA and the index measured in the JRC-EF task database were tightly correlated across certain types of occupations such as intellectual tasks, information processing and analytical tasks, but less tightly correlated for physical tasks. All in all, Mr. Sostero and co-authors’ work found that OJA data emphasize intellectual and social tasks, but mention physical, non-digital tools much less frequently, thereby representing work through the lens of the professional class.

Daniel Samaan (ILO Research Department)

Mr. Samaan’s presentation “Using Data Science to Analyze Labour Markets- New Insights on Job-to-Job Transitions” focused on i) how data science techniques can be used to analyze labour markets, and ii) how these data can be made part of labour market information systems. The presentation was based on a research brief, “Who moves and who stays? Labour market transitions under automation and health-related restrictions”, written in collaboration with co-authors at ILO and eightfold.ai. Mr. Samaan first clarifies that job-to-job transitions can be of many different types, such as by employment status, by occupation, within or across organizations among others. Given that job-to-job transitions are a central component of labour market dynamics, policymakers have an interest in providing frameworks that expedite transitions, as long transition spells can impose costs. Economies that can facilitate such transitions are more agile and flexible. Mr. Samaan mentioned that data sources on transitions are indeed limited, although experts expect transitions to increase in the future, and that people will change their career paths more than in the past. Given the lack of good data, Mr. Samaan and co-authors reached out to eightfold.ai (who have a database of approximately 1 billion CVs), and were able to create an occupational mobility network with this CV data from eightfold. Through natural language processing (NLP), AI technology, and a subset of CVs from eightfold (only American data), the authors were able to extract relevant data, i.e. data on people moving from one job to another, as well as other socio-economic data. There were differences between occupations in the number of transitions occurred. Of 674 occupations present, the machine learning algorithm proposed 28 “communities”, or groupings of occupations. Within specific communities, job transitions happened more frequently than across communities. Comparing these 28 communities, the authors find that most transitions happened in the “General Business and Management” community. Other highly specialized and highly skilled communities (like medical doctors and health specialists) presented less transition possibilities.

Veronica Escudero and Hannah Liepmann (ILO Research Department)

The presentation jointly presented by Ms. Escudero and Ms. Liepmann was on the topic of “Using online vacancy and applicants’ data to study skills dynamics”. Ms. Escudero started the presentation by noting that there exists a rich literature on skills dynamics in Europe and the US, owing to data from O*NET. Findings from US data suggest that computer technology replaces routine work, although interactive skills have been found to be less replaceable. However, there is a lack of knowledge and data on aspects of skills dynamics outside of high-income countries. The authors of this study make use of an alternative approach by using big data and online labour platform data to measure skills dynamics. This approach has the benefit of being able to measure changes in skills both within occupation, as well over time due to the panel structure. The objective of the study is therefore to see whether skills dynamics can be studied using online vacancies as well as job applicants’ data for developing countries. The authors develop a conceptual taxonomy based on the social science literature applicable to developing and emerging economies, and then implement this taxonomy using data from the Uruguayan portal BuscǎJobs, which contains information both on supply and demand. Borrowing from the literature in labour economics and psychology, the authors construct 3 broad categories - cognitive,
socioemotional, manual skills and 14 sub-categories, allowing for a more granular analysis. These categories are based on keywords and expressions (using machine learning and NLP), referred to as “unique skills”, which are then matched to vacancies or the applicant job spell. The BuscoJobs data contains information from 2010-2020, and data on job vacancies, applicants’ work biographies as well as their search behavior. A caveat is that the data are not fully representative, given that applicants tend to be younger and more educated relative to the workforce. One finding is that clerical support workers are overrepresented in the BuscoJobs data relative to the Uruguayan household survey, although there is no change of such bias over time. The usage of machine learning and NLP techniques allow a classification of more than 73% of applicants’ observations and 91% of vacancies. Finally, the authors were able to create 2-digit ISCO-08 occupations for both the vacancies and job applicants’ data. This allows an assessment of how skills can change over time within occupations.

Q&A session

The first question was directed towards Mr. Matteo Sostero (by Mr. Tony Bonen) about the JRC-Eurofound Task Database, regarding the source of the tasks. Mr. Sostero replied that there are multiple databases used. The reference data is based on a combination of the European Working Conditions Survey and the Survey of Adult Skills (PIAAC), as well as ad-hoc calculations from the European Union Labour Force Survey (EU-LFS). Mr. Bonen further stated that the finding that the profile of job posting candidates was not overly skewed towards professional/white-collar jobs was also obtained in the Canadian data.

Ms. Strietska-Ilina then alluded to a question raised by Mr. Delautre regarding how calculations in job-to-job transitions raised in Mr. Samaan’s presentation could be linked to policy conclusions as regards re-skilling and up-skilling. Mr. Samaan replied he did not speak about education and skills, as the team had not used this in the creation of the network. He did specify, however, that it would be very interesting to see if educational transitions went hand in hand with job-to-job transitions, as in principle, this information could be extracted from the CVs. Mr. Ekkehard Ernst further commented that ‘seniority profiles’ (seniority which people obtain over time), which Eightfold.ai possess, could be used to see if candidates’ current job possibilities depend on previous experiences.

Mr. David Ticell (LMIC) raised the point that much of the data raised in the discussions originates from private sector sources, some of which may not be very transparent. He asked what risks/benefits do participants see from this increasing reliance on private data and how to mitigate these risks? Mr. Sostero acknowledged this as an important question. He noted that in the EU, there has been an attempt to collect this data publicly. In particular, Cedefop (European Center for the Development of Vocational Training) has been working for a number of years to collect and make data on online job advertisements public. One risk that Mr. Sostero brought up was that there could be a potential bias in terms of skills considered, as they might be more geared towards professional jobs, and therefore not be representative. Ms. Escudero agreed with Mr. Sostero’s points, but also added that it is important to distinguish between different types of data. For instance, labour intermediation platforms are private by definition. Mr. Picell further identified a lack of openness and high costs of the data as potential risks as well.

Mr. Alfredo Alvarado asked a question of all speakers regarding the role of new technology adoption on skills requirements in developing technologies, especially in the context of labour-saving technologies. Ms. Escudero referred to literature using survey data, which finds that in emerging and developing countries, routine work is more strongly used than in advanced countries. Consequently, such aspects of technologies might affect developing economies lesser. Ms. Liepmann referenced a seminal study on the impact of technology used in Africa, which found positive employment effects across the qualification distribution. Ms. Strietska-Ilina alluded to broader research on skills and employability, and noted that what makes an individual successful in the labour market is a combination of foundational skills, technical skills and cognitive skills.
Finally, Mr. Aurelio Parisotto wanted clarification on how the algorithms can be tested for validity. Mr. Samaan answered that in some cases, it is hard to test for representativeness of the sample of online job postings, as a picture of underlying true population breakdown of the labour force is unknown. However, for the United States, official information on job-to-job transitions and occupational data is available. Consequently, data obtained from Eightfold.ai can be compared to official statistics. In the case of the United States, this bias does not seem to be do problematic, except in some cases (such as management occupations) which are overrepresented, and some which are underrepresented. Ms. Escudero mentioned that in her study, the authors have access to vacancy data as well as applicants’ job histories. The former is compared to LFS, whereas the latter is compared to surveys of enterprises to see job openings in the market. Mr. Bonen mentioned that samples could be adjusted based on observed biases relative to official statistics at a higher level of categorization, although that would be also subject to assumptions. Mr. Sostero noticed that the lack of observability of all posted jobs (including those not posted online) is at the core of the sample selection issue. In other words, jobs posted online might be systematically different in important ways to those that are not in ways that can bias analysis,
Session 2:

Introductory remarks

Opening remarks by Aurelio Parisotto (Employment Policy Department, ILO)

Mr. Parisotto, the moderator of this session, started by noting that this session was more macro in nature, relative to the previous one. Specifically, this session focused on how machine learning and data science techniques could be used to learn more about the macro determinants of unemployment, sectoral output, economic activities and labour mobilities. This session was comprised of 3 speakers.

Presentations

Nicholas Denis (Statistics Canada)

Mr. Davis’ presentation, entitled “In-season crop area estimation using satellite imagery”, with co-authors, provided an overview of how machine learning techniques and data science were being used at Statistics Canada, along with a practical example. Mr. Denis began the presentation by noting that data science (DS) brings value to policymakers at various levels of government in two ways. First, it allows for the creation of high-quality data products which can be used by policymakers to formulate policy. Second, DS is also used to deliver services more efficiently by improving upon previous products, by reducing the response burden in surveys for example and automating manual work. Mr. Denis shared several examples of how data science is used at Statistics Canada, including image processing, NLP, predictive analytics, data engineering and pipeline among others. Some types of data used at Statistics Canada include retail points-of-sales (scanner data), international trade transactions, cargo manifests, financial statements and more.

The example shared in the presentation by Mr. Denis was that of satellite imagery, with an application to in-season crop area estimation. Mr. Denis stated that every crop season, Statistics Canada produces a series of sampling surveys. In this survey, at the beginning of the season, crop producers are sampled and asked their intention to seed, including seed area. Further along in the season, they are surveyed again, this time asking how much area was actually seeded. Later in the season, they are asked their estimated yield, and at the end of the season, they are asked to provide their actual yield. There are two issues with this survey: only a small number of producers are sampled, and this is a costly endeavour. One product that Statistics Canada has been using is produced by Agriculture and Agri-Food Canada, who produce a land cover classification map of Canada every year, which classifies land into different types such as wheat, canola, land, body of water etc. into pixels. This endeavour can take 6-8 months to produce, and is not therefore timely. This is where Statistics Canada provides value, through the Ag-Zero initiative, which include zero-contact statistical products. The Ag-Zero project leverages freely available satellite imagery (Landsat-8), which has a periodicity of 2 weeks or less, with a coverage of crop fields of more than 98%. These images go through a machine learning model, get processed, and then throughout the crop-season, new estimates of crop area are produced. A Bayesian neural network model is trained on the data, and a measure of uncertainty is obtained to adjudge whether the estimates are reliable. At the moment, the project involves 46 crop classes, for the provinces of Alberta, Manitoba and Saskatchewan, and for the years 2014-2020.

At the end of the presentation, Mr. Parisotto noted that some participants asked about the costliness of the techniques used, and how feasible would it be to use such techniques in poorer countries.
Matthieu Charpe (Employment Policy Department, ILO)

Mr. Charpe’s presentation, entitled “Night lights as a proxy for economic growth? An application to structural transformation”, was a two-part presentation. In the first part, Mr. Charpe discussed how satellite images can be used to proxy economic growth, especially in countries where official statistics are lacking. The second part dealt with an application of this concept. Of the various data challenges in medium and low-income countries, some include the presence of a large informal economy, unreliable price indices and underfunded statistical agencies. Consequently, it is difficult to monitor the economy in real time, and therefore gauge the impact of policies and/or events. A proposed solution is to use satellite images at night, or nighttime lights (NTLs), which make use of data collected at the earth’s surface using remote sensing. The main idea is that lights observed from space can be used to gauge the level of human activity, which serves as a proxy for economic activity. One challenge from using such data is that it involves heavy data processing, as the raw data captures human activity as well as non-human activity, the latter of which must be separated out. The Visible Infrared Imaging Radiometer Suite (VIIRS) captures this data monthly, and the measure of light captured is at the pixel level (at 1 km² at the equator) for every location on the earth, except at the poles. This data can then be aggregated to any level, such as the city, country, or regional level. Existing literature has found correlations (elasticity) between light activity and GDP for the world, as well as different income groupings of countries. An important limitation of this approach is that components of GDP such as services and manufacturing are more intensive than agriculture, and therefore, light activity might fail to capture economic activity in regions which are agriculture-intensive. Moreover, it could be hard to disentangle population effects from income effects as well.

Mr. Charpe then related this literature to the work done at the ILO, and asked if there is actually a light-employment elasticity. Two approaches can be used to answer this question. The first could use Okun’s law to back out the light-employment elasticity from the light-GDP elasticity. A second makes use of data available at the ILO (national level), microdata repository (regional level) and local data (census level). Applications of NTLs include revisiting papers which make use of GDP measures and replacing them with light intensity, Covid-19 using NTLs in India and Morocco, climate change, conflicts, regional inequalities, nowcasting of employment, how minimum wages can affect light activity (and unemployment), among others. Mr. Charpe then presented another application which he has been working on, and which asks if poorer areas are catching up with richer areas in sub-Saharan Africa at the local level. This application makes use of light per employment as a proxy for local labour productivity, and finds that there is indeed convergence at 2% annually, but there is heterogeneity as some areas are left behind.

At the end of Mr. Charpe’s presentation, Mr. Parisotto also noted that this methodology might be biased against agriculture, which is not light intensive as other activities.

Linus Bengtsson and Tom Smallwood (Flowminder.org)

The third and final presentation of the session was co-presented by Linus Bengtsson and Tom Smallwood, of Flowminder.org, on “Estimating population mobility from mobile phone usage data”. This presentation sought to understand population mobility, migration and dynamics of movements in low and lower-middle countries. Mr. Bengtsson began by noting that there are a lot of dynamics in the data that can also potentially be related to the labour market. As a bit of a background, Mr. Bengtsson noted that Flowminder.org is a non-profit organization, which works with many governments and international organizations, and was launched after the earthquake in 2010 in Haiti and the subsequent displacement. Mr. Bengtsson then yielded the floor to Mr. Smallwood, who then gave a description of Call Detail Records (CDRs), which are generated every time cell phone users make or receive phone calls, text messages or use mobile data. CDRs can give insight into the large-scale
movement patterns of the population in lower middle-income countries. As CDRs are generated automatically, they allow the data scientist to gather data in near real-time. One concern related to using CDR data is that the population captured by this data is not necessarily representative of the underlying population. In particular, being in the dataset can be affected by factors such as age, gender and socio-economic status. Mr. Smallwood then described some of the key principles adhered to by Flowminder.org, including privacy and transparency of individual subscribers. This is done by complying with GDPR, as well as the appropriate country legislation in which it is operating. Mr. Smallwood highlighted that the use of CDR data in countries involves collaboration with various stakeholders, such as mobile network operators (MNOs), government end users such as national statistical officers, regulators, as well as intergovernmental organizations. CDR data allows for a broad range of mobility indices to be derived, such as crowdedness, population mixing, where people are located over time, inter and intra-regional travel, among others. As an application of CDR data, Mr. Smallwood discussed work done by Flowminder.org in Haiti on how the location of subscribers changed between day and night, thereby allowing identification of important commercial and residential areas, and patterns in daily movements of people. As another application. Mr. Smallwood alluded to work done in Namibia, which looked at seasonal changes in population living in different areas, which in turn potentially reflected labour migration based on specific sectors such as agriculture and tourism. Similar work was done to map monthly migration flows in the Democratic Republic of Congo. Mr. Smallwood further highlighted the role that CDR could play in analysing mobility disruption during a crisis. Lastly, Mr. Smallwood described several ways in which mobile phone data usage could be used to profile subscribers. For instance, airtime purchases, recharge amounts, spending speed and such could give insight about the financial profile of a user. Mobility then could be spliced by different subgroups of the population, based on such profiling.

Q&A session

Mr. Mauricio Gallardo (Universidad Católica del Norte, Chile) asked two questions directed to Mr. Denis. The first question was about the measurement accuracy of the satellite imagery data relative to traditional methods which are used by national statistical offices. The second question noted that there are other opportunities to use such data, such as measuring economic activity in the construction sector. Bearing that in mind, he then asked if Mr. Denis saw potential uses such as this. Mr. Gallardo then asked a question of Mr. Charpe, and the potential introduction of biases due to the inherent differences in the intensity of certain daytime tasks compared to nighttime tasks. For instance, leisure activities might be more intensive at night, while agriculture and construction might be more intensive during the day.

In response to the first question, Mr. Denis reiterated that the inputs in the model are satellite imagery of a specific field, which are labelled using administrative data (crop insurance) at Statistics Canada, at the provincial level. As a measure of model performance, the model outputs are compared to the ground truth at the _aggregate level_. Regarding the second question, Mr. Denis acknowledged that there are projects which measure construction activity in Canada, which capture images at different stages of construction at particular sites.

In addition to the question asked by Mr. Gallardo, there were further questions about the usage of NTL methods in India, and if the methods could be informative about the informal sector. Mr. Charpe replied that there are studies which look at construction activity as a proxy for income growth, but these are plagued with difficulties. Moreover, NTL can be thought of being more informative about income growth, rather than levels, given that there may be systematic country level differences in the level of light use at night. To the question regarding informality, Mr. Charpe noted that NTL are a measure of electricity consumption, and they are therefore able to capture the informal sector better. Finally, Mr. Charpe reference a paper on regional inequalities, as well as poverty and secondary cities based on NTL in India. He further referenced another study published in _World Development_, which looked at electricity consumption and nighttime light in the context of Covid-19 in India.
Mr. Parisotto then posed another question, which asked if such methodologies could be used for progress in the SDGs. Mr. Bengtsson referred to a UN working group using big data for official statistics to this effect.

Mr. Parisotto raised a further issue, regarding the usage of satellite images, and if this work can be applied to developing countries, where economic activity may depend predominantly on agricultural production. Mr. Denis replied that there exist gradations of cost. For instance, Statistics Canada builds these models from scratch, and although satellite image data are free, they are at a lower resolution. To make these models, however, ground truth labels are also required, which may or may not be available for free. Moreover, constructing these models would also require good hardware (ex. GPUs), which if done remotely (i.e. via cloud) could cost about $100 daily. However, once the model is built, using it is cheap.

Finally, Mr. Delautre asked if Flowminder.org would want to expand to more countries, to which Mr. Bengtsson replied in the affirmative. He referred to a platform which is being built by Flowminder.org, that produces aggregated data, and which will be available for free for Haiti. He expressed his hope that this will be replicated in other countries.