MIGRATION POLICY CENTRE

Funded by The European Union

GS4S
Global strategy for skills migration and development

GS4S Working paper series (D7.3)
Working paper no. 3

Locating shortages in migrants' origin countries: a big data approach

This working paper is part of the Horizon Europe project GS4S - Global Strategy for Skills, Migration and Development (gs4s.eu).
Project deliverable: D7.3 in T3.1

Author: Friedrich Poeschel

Reviewers: Tommaso Frattini, Mahdi Ghodsi, Pascal Beckers

Locating shortages in migrants' origin countries: a big data approach

Friedrich Poeschel*

Abstract: This note documents a data collection on vacancies published online, which is being implemented by web scraping online platforms in selected non-EU countries. The data collection aims at locating labour or skill shortages in important origin countries of migration to the EU. Where the shortages coincide with shortages in EU countries, a skill partnership could address both shortages simultaneously. The potential of the web scraped data are explored based on the initial wave of the data collection. The note concludes by outlining how the data collection can be transformed into measures of shortages at the level of occupations and skills.

Keywords: shortages, online job adverts, web scraping, vacancies, BLMA, global skills partnerships

Acknowledgement: This paper is part of the Horizon Europe project GS4S - Global Strategy for Skills, Migration and Development (gs4s.eu). The funding from the European Union is gratefully acknowledged. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, Horizon Europe or the Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Helpful comments and suggestions were received from colleagues in the GS4S project, its Social and Scientific Advisory Board, and participants of the annual meeting of the European Network on Regional Labour Market Monitoring in Lugano.

Notes on Kosovo & Palestine: Kosovo - This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo Declaration of Independence. Palestine - This designation shall not be construed as recognition of a State of Palestine and is without prejudice to the individual position of the Member States on this issue.

* European University Institute. Migration Policy Centre, Robert Schuman Centre for Advanced Studies, European University Institute. Via delle Fontanelle 19, 50014 Fiesole, Italy: friedrich.poeschel@eui.eu.

Funded by
The European Union

Contents

Funded by
The European Union

1. Introduction

The basic logic of global skills partnerships is rather compelling: when both an origin country of migration and a destination country face labour or skill shortages in the same occupation, they can both gain from training in the origin country that targets both countries' shortages. Such a common interest in addressing shortages can serve as basis for a global skills partnership between two countries. While many bilateral labour migration agreements (BLMA) might not require a shortage on both sides in order to function and receive government approval, either side may be more interested in creating a scheme when it could address a pressing domestic issue. Whenever this requires significant investment in training facilities – which may only become possible through the partnership – a sufficiently long time horizon is needed for such investments to pay off. A common interest to maintain the partnership will likely be a stabilising factor, in contrast to partnerships that are primarily in the interest of only one country.

A migration partnership that relies on shortages in both the destination and the origin country therefore needs at least rudimentary information on which occupations or skills are in shortage, and the data collection presented in this note is meant to contribute to this objective. For destination countries in the European Union (EU), a range of shortage indicators are typically available, such as views of employers on hiring difficulties (for example in the Investment Survey of the European Investment Bank). However, this example of subjective views also highlights that an exact science of identifying labour or skill shortages does not exist. Various indicators have been used both in academic research and in policy-oriented reports, thus far without reaching a consensus on which indicator is most suitable. That said, shortage indicators often use data on job vacancies in one form or another (van Smoorenburg, 2024). Apparently, job vacancy rates are the only shortage indicator that is systematically available for almost all EU countries on Eurostat (notably the quarterly time series jvs_q_nace2). Job vacancy rates are obtained by relating the total number of current vacancies to current total employment. In academic research on labour economics, the concept of 'market tightness' is used very widely (Duval et al., 2022), which relates the number of current vacancies to the number of currently unemployed persons.

The data on job vacancies themselves may be collected through dedicated employer surveys (such as the IAB Job Vacancy Survey) or may be based on information from public employment services, with the caveat that many vacancies – perhaps even the majority – are never registered with the public employment service (see e.g. Antolín, 1994). However, such data sources on vacancies are apparently not yet commonly available for origin countries of migration from outside the EU. This may be because data on vacancies are simply not collected, or data are collected but are not made available to third parties, and even if they are available in principle, obtaining them may be very tedious in practice.

When shortages in origin countries remain invisible due to lack of data, policy makers in destination countries cannot know which occupations to suggest for mutually beneficial skills partnerships, key stakeholders in origin countries might not know which suggestions are in their interest, and evaluators of skill partnerships might find their task impossible. This paper therefore

Funded by
The European Union

adopts another, big-data method of collecting vacancy data for non-EU countries: web scraping the information that is freely available from job vacancy adverts published online (henceforth OJA for online job adverts). To the extent that information from OJA can be translated into policy-relevant shortage indicators, they could fill this gap in the available data for origin countries outside the EU.

The paper proceeds as follows. Section 2 reviews the literature on OJA, with a focus on recent contributions and with some examples for their use in research and policy-making. Section 3 specifies the selected sample of non-EU countries and the targeted websites in each country, as well as the way the websites were chosen. The web scraping method applied to these websites is described in Section 4. Section 5 explains the layout of the scraped data so far and makes a first assessment of data quality. In light of this assessment, Section 6 discusses how policy-relevant shortage indicators may eventually be derived from the data.

2. Literature review

Online job adverts (OJA) have become a very popular data source for vacancies in research and analysis related to labour markets; examples in academic research include Marinescu and Wolthoff (2020) and Yeh et al. (2022). But also recent empirical work addressing a wider audience – including policy makers, social partners, and the public at large – draws on OJA data, e.g. OECD (2022), CENTAR's Gigmetar and Bertelsmann's Jobmonitor. Much of this work emerged only in recent years, as data from large-scale web scraping became increasingly available. With regards to migration studies, however, Tjaden (2023) finds that web scraping has so far remained "underused", while potentially very promising for analytical insights that are hard to obtain otherwise.

Like other data on vacancies, OJA data have also been used to measure the demand for labour or certain skills (Acemoglu et al., 2022) and to detect shortages. A frequent measure for labour demand and shortages are job vacancy rates, relating vacancies to existing employment in the sector or occupation considered. This concept was already used in early analyses of vacancies, such as Jackman et al. (1989), and in the field of public health (e.g. Batata, 2005). A recent quantitative evaluation of various shortage measures by van Smoorenburg (2024) found that job vacancy rates perform well indeed, while market tightness (unemployed job seekers relative to vacancies) is found to be the best performer. Also the change in vacancy rates has regularly received attention (e.g. in the work by Teo et al., 2022 on nurses). After all, a *high* vacancy rate might reflect high turnover of employees in a specific sector or occupation, but then a *rising* job vacancy rate would still provide important clues.

A significant part of the literature on OJA revolves around issues of data quality. A recent contribution by Plaimauer (2024) provides an overview of important issues, which will also be discussed as part of a first quality assessment in Section 5. A key problem is that one OJA does not necessarily correspond to one actual vacancy: on the one hand, original adverts are replicated on other websites and outdated adverts can stay online, which tends to inflate OJA numbers beyond actual vacancies. On the other hand, a single OJA can be used to promote several vacancies (and if so, this is often unobserved), which tends to deflate OJA numbers relative to actual vacancies.

Funded by
The European Union

Despite the concerns over the quality of OJA data, they may well be the best available data source on vacancies – given the limitations of employer surveys and the subset of vacancies registered with the public employment service. Beręsewicz et al. (2021) conclude that the Polish vacancy survey among employers also does not reach full coverage, mainly because employers do not always respond and sometimes under-report their vacancies. It has also been suggested that non-response is more frequent for SMEs, as larger firms have more resources to deal with surveys. Meanwhile, evidence from Ghana suggests that vacancies not published online are also harder to find for job seekers (Lambon-Quayefio et al., 2024). OJA may also be especially useful for disaggregated analyses such as for regions or specific occupations (Vermeulen and Gutierrez Amaros, 2024). The full text of the OJA allows for rich further analyses, e.g. deriving information on non-wage amenities (Escudero et al., 2024).

A different but related question is whether the issues with OJA data undermine their value for empirical analyses. That is, even if OJA over- or underestimate actual vacancies as explained above, the behaviour of actual vacancies might still be inferred from OJA to a reasonable degree of accuracy. Indeed, Evans et al. (2023) find that OJA combined with a suitable algorithm can capture the evolution of actual vacancies. Similarly, OJA data may still be useful for locating shortages, as is the objective in this paper. For example, a recent contribution by Brown et al. (2024) finds that certain aspects of OJA can predict employers' assessments of shortages, as collected in employer surveys.

3. Targeted online job adverts

The possible use of this work as input for skills partnerships has determined the geographical focus: countries outside but relatively close to the EU, where there is (potential for) labour migration to the EU and some form of co-operation or mutual dependency already exists. Concretely, the 28 partner countries[1] of the European Training Foundation (ETF) arguably capture most of these countries. In several cases, these countries have already entered in bilateral or multilateral co-operation with EU Member States (e.g. Germany's Westbalkan-Regelung). Their partnership with the ETF means that further insights and statistics related to the labour markets of these countries are available, which could be used to complement and cross-validate the OJA data collected in this project and could more generally help prepare skills partnerships with these countries.

[1] In alphabetical order: Albania, Algeria, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Egypt, Georgia, Israel, Jordan, Kazakhstan, Kosovo (see note on front page), Kyrgyzstan, Lebanon. Libya, Moldova, Montenegro, Morocco, North Macedonia, Palestine (see note on front page), Serbia, Syria, Tajikistan, Tunisia, Türkiye, Turkmenistan, Ukraine and Uzbekistan, as indicated by the ETF on https://www.etf.europa.eu/en/where-we-work .

Funded by
The European Union

Table 1: ETF partner countries prioritised in the data collection

country	code	total population 2021	total EU residence permits received in 2022	of which: employment-related
Albania	ALB	2,811,666	79,169	21,052
Algeria	DZA	44,177,969	42,121	4,551
Belarus	BLR	9,340,314	310,285	145,788
Bosnia and Herzegovina	BIH	3,270,943	50,221	30,265
Egypt	EGY	109,262,178	37,698	8,790
Georgia	GEO	3,708,610	38,042	25,228
Kosovo*	XKX	1,786,038	46,879	19,808
Libya	LBY	6,735,277	3,472	238
Moldova	MDA	2,615,199	29,492	13,631
Montenegro	MNE	619,211	4,119	1,935
Morocco	MAR	37,076,584	162,779	43,824
North Macedonia	MKD	2,065,092	29,591	11,623
Serbia	SRB	6,834,326	52,842	26,684
Tunisia	TUN	12,262,946	43,014	13,197
Türkiye	TUR	84,775,404	120,270	39,522
Ukraine	UKR	43,792,855	382,401	275,719

*This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo Declaration of Independence. Sources: IBAN (country codes), World Bank (population figures), and the Eurostat residence permit data collection (mig_res).

Within the set of the 28 ETF partner countries, the OJA data collection prioritises those in close proximity to the EU, notably countries in the Western Balkans, the EU Eastern Partnership countries and North African countries, as listed in Table 1. However, the aim is to eventually cover all or almost all ETF partner countries. The residence permit figures reported in Table 1 show that many of the selected countries are already important origin countries for legal migration to EU countries, including employment-related migration. This is not limited to the countries with larger populations such as Türkiye and Morocco but also includes relatively small countries such as Bosnia and Herzegovina, Kosovo[2] and Moldova.

The data collection necessarily has to set priorities also with regard to the websites covered for each selected country. The focus is on open platforms where a variety of OJA may be posted and at least 70-100 adverts are currently listed. Websites that cover only jobs in a few specific occupations, a particular sector or a particular kind of firm (e.g. consultancies) are left out for now as they typically do not meet either criterion, being neither cross-cutting nor listing sufficiently many adverts. The most well-known international platforms such as monster.com as well as LinkedIn and some platforms of public employment services have also been de-prioritised for the time being: these sites are typically programmed and operated with more advanced technology.

[2] This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo Declaration of Independence.

Funded by
The European Union

At best, a more complex web scraping method will be required, but these sites may often altogether block any attempts of web scraping.

Table 2: Websites used in the data collection thus far

Albania	Kosovo*	North Macedonia	Serbia
gjirafa.com	gjirafa.com	gjirafa.com	halooglasi.com
njoftime.al	kosovajob.com	kariera.mk	lakodoposla.com
njoftimefalas.com	lyppune.com	njoftimeperpune.com	kupujemprodajem.com
njoftime.com	merrjep.com	pazar3.mk	oglasi.rs
njoftimerperpune.com	ofertapune.net	vrabotuvanje.com	poslovi.infostud.com
njoftime365.com	portalpune.com		
	shpalljepune.com		
Algeria	**Bosnia and Herzegovina**	**Moldova**	**Montenegro**
algerie.tanqeeb.com	mojposao.ba	angajat.md	prekoveze.me
bayt.com	szzhnz-k.ba	bestjobs.eu	zaposli.me
emploipartner.com	szztk.ba	md.trud.com	zzzcg.me
naukrigulf.com	zzzrs.net	rabota.md	

Note: As some websites utimately turned out to be inaccessible (especially in Albania, Bosnia and Herzegovina as well as Kosovo*), not all websites in the table are covered by the database.
*This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo Declaration of Independence.

Table 2 lists the websites for the countries covered so far by the data collection. These websites were found by googling with search terms such as “job offer”, “open position” and “vacancy” in the language(s) of the targeted country, noting that several countries in Table 2 have more than one official language. When googling, sometimes further search terms were identified from the first search results and were then tried in turn. In addition, colleagues who originate from the targeted countries were asked which search terms they would use in this context. Once some suitable websites had been found, *similarweb.com* was used to identify further such websites, typically pointing to platforms of OJA in the same country or the same language. With this approach, the search results on Google relatively quickly converged to a limited number of websites, which is not surprising: the business model of such platforms relies on attracting traffic, and they engage in search engine optimisation of their websites to ensure that they are easily found on Google.

4. Web scraping method used

Platforms of OJA are normally written in the programming language HTML (for Hyper Text Markup Language), which effectively structures the blank space of a website and determines where each element of the website is placed, such as text and images. At the same time, HTML stores the attributes of each element, such as colour, size, font or a hyperlink. As a result, the code underlying a website quickly becomes unwieldy. However, a related programming language called

Funded by
The European Union

XML (for Extensible Markup Language) can be used to find and select specific elements in the HTML code, precisely by drawing on the structure and the attributes in HTML to construct an XPath. This allows for automatic queries directed at websites: once contacted successfully, the HTML code of a website is loaded, and by applying the prepared XPaths, specific elements (or attributes) are filtered out.

The key challenge in this context is to ensure that the XPath retrieves the desired data and only the desired data. An even slightly misspecified XPath will retrieve nothing or the wrong data, while a seemingly correctly specified XPath can retrieve too much – either lumping together several instances of the desired data or retrieving a collection of data in which the desired data are only a small part. For example, when the desired data appear at varying places among "siblings" (i.e. at the same level of the parent-child hierarchy in the HTML code), it can be hard to filter out from these siblings, and one has to come up with an XPath that somehow still uniquely identifies the desired data. In practice, this means that web scraping requires learning the XML programming language, which was an important preparatory step in this project.

As software solution to implement the web scraping procedure – i.e. to rapidly contact websites, apply the XPaths, and store the retrieved data – various options were considered, including the well-known solutions Python and R. Ultimately, the KNIME analytics platform was chosen as the most suitable for this data collection, because

- it is a free open-source software that nevertheless enjoys professional support (including regular updates) by a company headquartered in Zurich, after initial development at the University of Konstanz in Germany (KNIME = Konstanz Information Miner)
- its open-source nature means that new or improved methods in data science become available on KNIME relatively quickly, so that existing programmes can be adapted or upgraded
- its modular approach (not based on individual lines of code but on combining packages of code) makes the programming more stable, transparent and replicable, and thereby more accessible for other users.

The web scraping procedure that was purpose-built in KNIME for this data collection has two stages, corresponding to two instances of web scraping required for every platform of OJA considered so far. The first stage works on the HTML code of overview pages – those pages that display only "teaser" information for a number of OJA – and scrapes the hyperlinks to the dedicated page for each OJA, where the full information is displayed. The second stage works on the HTML code of the dedicated pages one at a time, scraping information such as job title, location, employer and – whenever provided – information on the wage, sector, full-time vs part-time, expected level of education and expected prior work experience.[3] Both stages of the

[3] Thus far, the second stage has focused on the information displayed systematically in certain places of the dedicated page for the advert, not attempting to locate information in the "free entry" text of the advert.

Funded by
The European Union

procedure in KNIME are designed as loops over all overview pages and all dedicated pages, respectively. As final part of the procedure, the scraped information is cleaned and converted into Excel files.

Considerable time was invested in building a prototype procedure and testing its performance on a few platforms of OJA, essentially by trial and error. A revised and refined version of the prototype has since been applied to a steadily growing number of platforms, and it is expected that data collection from almost all targeted websites of 8-10 countries will have been completed by the end of September. Some targeted websites turn out to be inaccessible, either because their HTML code differs due to an unknown manipulation of some conventions (a changed "namespace") or because they actively block IP addresses that send the kind of automised requests involved in web scraping. That apart, the progress of the data collection depends only on how quickly the prototype can be adapted to another platform of OJA (as each platform is coded differently, at least all XPaths have to be constructed anew) and on the speed of the internet connection. New XPaths are also required when platforms make substantial changes to their HTML code, which indeed occurred for one of the first platforms used in the collection.

5. Characterisation of the collected data

Information available. After the development and refinement of the web scraping programme, an initial wave of the data collection was assembled in August/September 2024. The websites used display a great heterogeneity in terms of the information they provide in an OJA. While virtually all platforms have some pre-defined fields for certain pieces of information, they vary strongly in which fields they define. All platforms provide the job title and the job location through such fields (unless the work is to be done remotely) but they vary widely with regards to further fields provided. Most provide a field indicating the employer, but in a significant number of cases, only some anonymous contact details are specified. Many platforms provide a field for wage information but in most cases, the field is left blank. Where wage information is provided, it may come in EUR, in USD or in other, local currency (and may be expressed as a range). Also a field on full-time vs part-time work is often included and if so, it is typically filled out. Another frequent field indicates (supposedly, see below) when the vacancy was posted.

By contrast, relatively few include some kind of sector information and if so, the categories used to group sectors are again very heterogeneous across platforms. Only some platforms include fields for the education level(s) or the prior work experience expected from applicants. Rather rarely, fields indicate whether the work is to be performed remotely and how many positions are available. However, in addition to the pre-defined fields, information is provided through the main text of the OJA, which can range from just something like "Those interested should call" to a detailed description of the employer, the job duties and the expected candidate profile. This description appears to be decided entirely by the user posting the OJA, as opposed to being somehow structured or standardised by the platform.

While this would be possible, it could also introduce frequent errors from misinterpreting the text. This is discussed further in Section 6.

Funded by
The European Union

On many platforms, users may even enter the information in different languages. The main text may be in English rather than the local language, and in a country with several large language groups, a substantial proportion of the OJA may be presented entirely in one of these languages (both the main text and the pre-defined fields). For example, some OJA on Moldovan platforms are presented in Russian. In general, the raw data are always collected and stored as they appeared on the original website; the website is not translated using built-in browser functions. Only in preparation of classification and evaluation of the data, they are translated, using deepl.com whenever possible (as only some languages are available there) and using Google Translate whenever necessary.

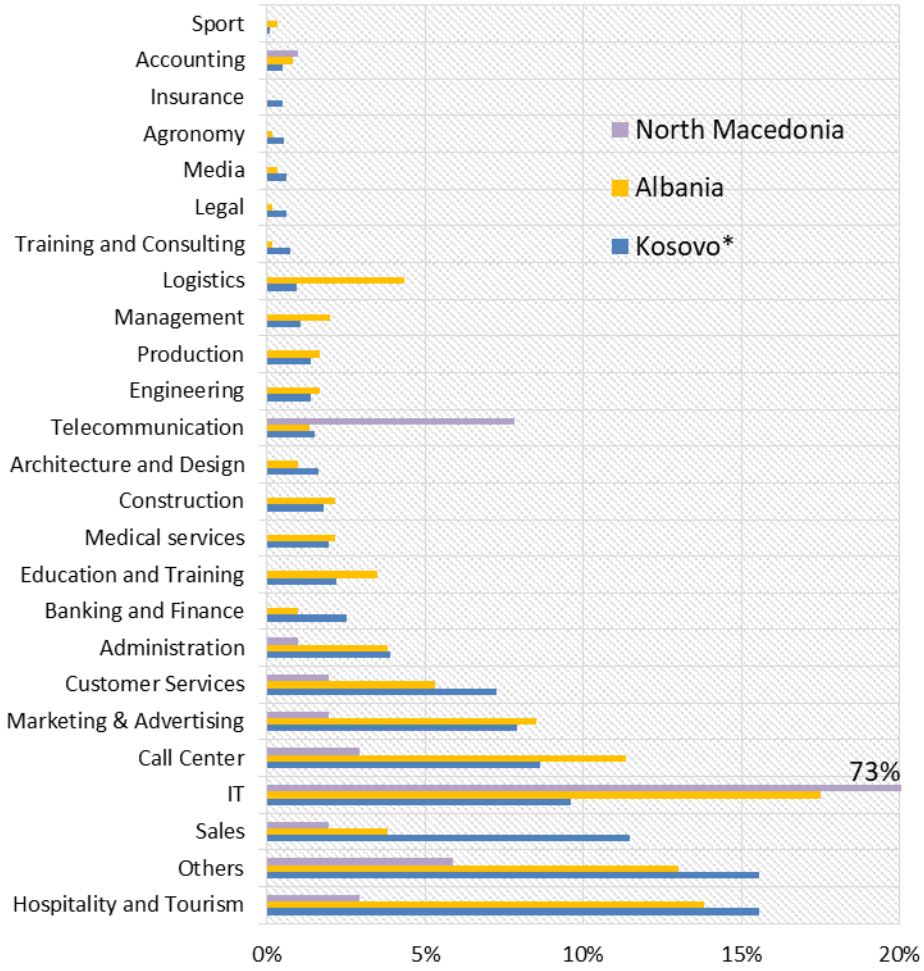
Sectors and occupations. Where information on sectors is available, it is not necessarily useful, as illustrated by the case of bayt.com for Algeria. Only eight categories are used to group sectors: "Hospitality & Accommodation" accounts for some 15%, "Management Consulting" for about 4%, "Sales Outsourcing" for about 3%, four further categories jointly for under 3% and "Other Business Support Services" for 75%. On the platform merrjep.com for Kosovo*, as another example, 15% of the OJA are categorised as construction, 5% as work from home, 4% as work abroad, 20% as other and 40% have no entry, while the remainder is spread out over 16 categories that each account for 3% or less. Also the sector information of other websites displays a tendency towards a limited number of categories, with ballooning categories for "other" or unspecified sectors. Not a single website can be reported to provide sector information using categories from internationally standardised classification systems.

Figure 1 shows arguably the most useful sector information encountered in the data collection so far, from gjirafa.com and covering OJA for Albania, Kosovo* and North Macedonia. Here the grouping results in a much more informative distribution at least for Albania and Kosovo*, with several categories accounting for substantial shares in excess of 10% of the OJA, several categories accounting for smaller but significant shares, and several categories accounting for insignificant shares. The category "Others" here accounts for at most 16%. In addition, the distributions for Albania and Kosovo* have a lot in common (while that for North Macedonia is strongly skewed towards a single category, possibly due to relatively few OJA for North Macedonia on gjirafa.com).

Such sector information allows for some first useful insights. In both Albania and Kosovo*, relatively many OJA concern jobs in IT, hotels and tourism, call centres, marketing and publicity, as well as customer services. IT is actually the most frequent sector posted in Albania (accounting for about 1 in 6 OJA) and North Macedonia (accounting for almost 3 in 4 OJA). In Kosovo*, jobs in sales appear to be posted especially often, compared to the two other countries. In North Macedonia, jobs in telecommunications and in Albania, jobs in logistics are posted especially often. Relatively few jobs are posted in medical services – no more than 3% in any country.

Funded by
The European Union

Figure 1: Sectors indicated on gjirafa.com (for three countries), Q3 2024

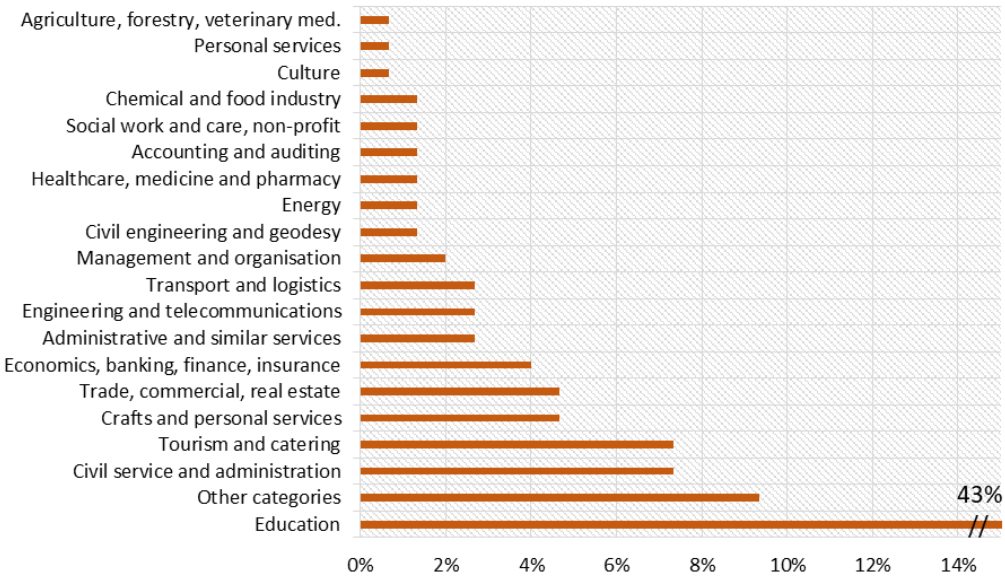


Note: Categories are translated only where the original designation in Albanian is not self-explanatory.
*This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo Declaration of Independence.

Figure 2 below shows the sector information included in OJA posted by a public agency of the Republika Srpska, one of two constituent parts of the Federation of Bosnia and Herzegovina. This figure can serve to illustrate which sectors might be under-represented on the private platforms for OJA, which could be the case in particular for the public sector and similarly for education and the health sector. Indeed, OJA for jobs in education appear to dominate on this site, accounting for almost 43% (64 of 150 OJA). Also civil service and administration (7%) is a relatively frequent sector, here at a par with tourism and catering. By contrast, jobs in healthcare, medicine and pharmacy account for less than 2% of the OJA, roughly in line with Figure 1. The IT sector appears to play only a minor role in Figure 2 but the true shares may be distorted, as the figure relies on relatively few OJA (150).

Figure 2: Sectors indicated on zzzrs.org (Bosnia and Herzegovina), Q3 2024

Funded by
The European Union



Even the comparatively good sector information shown in Figures 1 and 2 allows for only few comparisons across websites. In turn, incomparable categories from different websites cannot be aggregated to obtain a country-wide overview of OJA and derive insights on shortages. To overcome this problem, one would need to deconstruct the categories down to a level that is comparable across websites, and then aggregate them in the same way across websites. However, without more detailed sector information, this is not possible. While the vast majority of OJA indicate the employer, it is not clear how to infer the detailed sector the employer operates in. The company name can give clues, albeit often not detailed enough. A way forward may be business registries, which at least in some countries provide information on the sector(s) the employer operates in. In any case, the classification likely remains tricky.

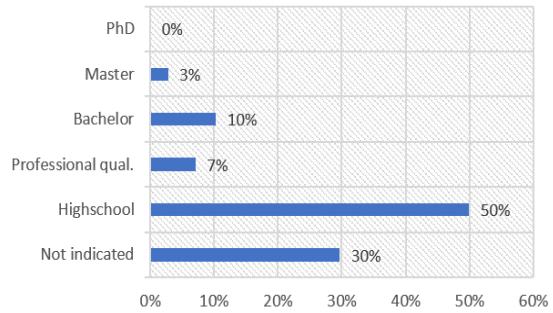
In sharp contrast, some kind of job title is available for each and every OJA, representing data at a very low level of aggregation (the level of the individual job). In all but rare cases, the job title will provide clues as to the occupation or at least some main task involved. As Section 6 will explain, this information may be used to infer exact occupations or detailed occupational groups. The OJA data may then be aggregated to occupational groups that are comparable across websites and even across countries, such as ISCO groups. In other words, while shortage indicators for sectors appear hard to reach, such indicators appear to be within reach for occupations.

Skill levels. The pre-defined fields encountered so far in the data collection include the level of education and the prior work experience expected from applicants. While neither necessarily matches the notion of skill level, both represent important aspects of skill levels. (Further information on skills can be found in the main text, see e.g. Gu and Zhong, 2023). Figure 3 shows the data on expected education that was obtained from websites for Albania and Moldova. The categorisation of education level appears comparable in a way that could allow for comparable aggregation across websites and countries, leading to relevant insights on labour demand and shortages. For example, Figure 3 suggests that around 20% of the OJA on both platforms require a higher degree, found as the sum of PhD, Master, Bachelor and professional qualification in Panel A and as "University" in Panel B. At the same time, just high school or literally any education level is typically acceptable for OJA on both platforms.

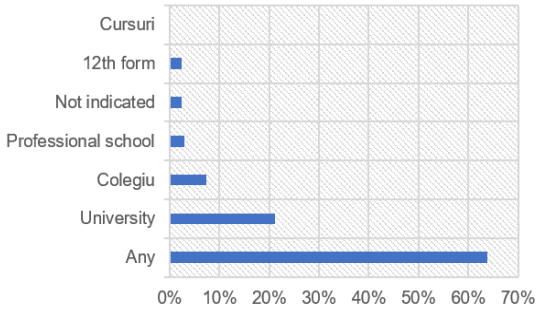
Funded by
The European Union

Figure 3: Information on expected education, Q3 2024

A. Education expected on njoftimefalas.com (Albania)



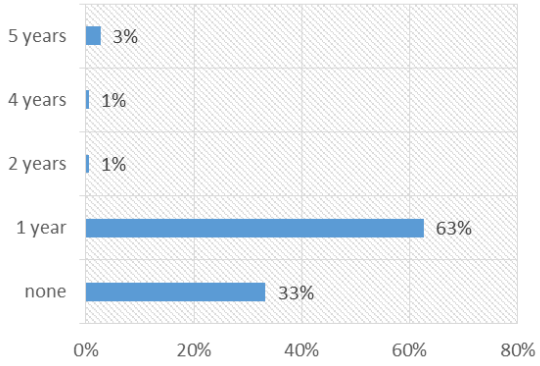
B. Education expected on rabota.md (Moldova)



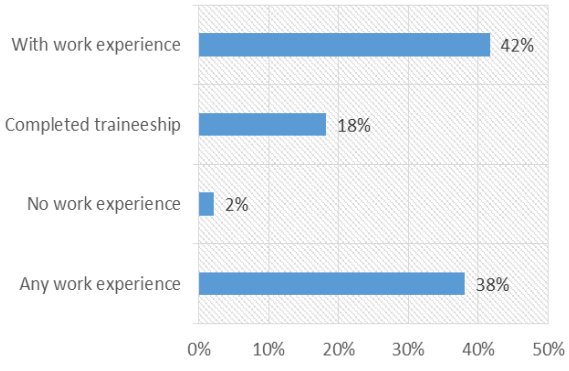
Regarding prior work experience expected from applicants, Figure 4 highlights the categories used in this context. On the platform catering to the Republika Srpska in Bosnia and Herzegovina (Panel A), a clear majority of the OJA (63%) indicate that at least 1 year of relevant experience is required, while very OJA require longer work experience (5%). As the remainder of the OJA does not require any work experience (33%), labour demand on this platform appears heavily skewed towards little or no experience, which together account for 96% of the OJA. On a platform from Montenegro with more than 1,000 OJA uses a different categorisation (Panel B). The jobs posted on this platform appear to require some relevant work experience about as often as no or any level work experience (40% each). However, nearly one in five OJA comes with a rather specific requirement - a completed traineeship.

Figure 4: Information on expected experience, Q3 2024

A. Experience expected on zzzrs.net (Bosnia and Herzegovina)



B. Experience expected on zzzcg.me (Montenegro)

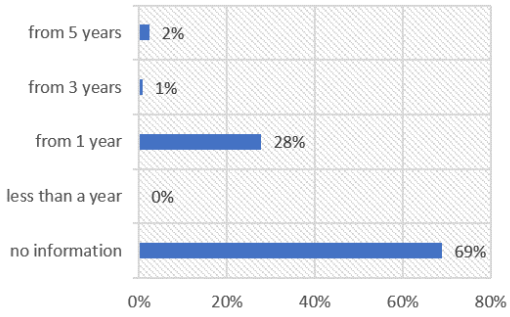


Funded by
The European Union

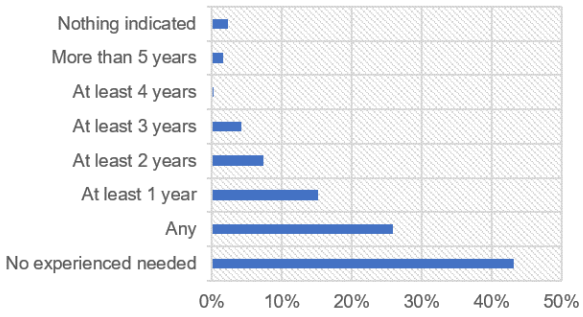
Figure 5 displays the data from two Moldovian platforms. This example therefore also serves to explore whether different sites in the same country suggest roughly the same conclusions (since the categories are again comparable). The two platforms are in agreement on the share of OJA that require at least one year of prior experience (around 30%). For the remainder, there is no information according to Panel A. In Panel B, roughly the same remainder is split between the entries for any experience and no experience needed. This suggests how the lack of information in Panel A may be interpreted.

Figure 5: Information on expected experience, Q3 2024

A. Experience expected on trud.com (Moldova)



B. Experience expected on rabota.md (Moldova)



Quality considerations. A number of challenges for data quality arise with these web scraped data. A key problem is duplicates: either an OJA for the same vacancy is intently posted on several platforms, or some platforms simply replicate OJA posted elsewhere. The latter phenomenon sometimes becomes visible when a platform indicates another platform as source for an OJA, but there may be many cases when this is not done – after all, it can make sense to copy OJA from other platforms in order to attract traffic for advertisements etc. However, it will be possible to weed out such duplicates by finding OJA that have the same job title, job location and employer; if still in doubt, also the main texts may be compared. In addition, only OJA that appear in the same quarter may be considered duplicates (while the exact date when the OJA was posted may vary), as employers commonly re-post the same OJA as before when another recruitment has to be made.

Avoiding duplicates could considerably reduce the more general aforementioned problem that OJA are not the same as vacancies. As another part of the same problem, more than one vacancy may be behind a single OJA. Since the Moldovian platform bestjobs.eu seems to require information on the number of positions (in a pre-defined field), the extent of this issue can be investigated using the collected data (168 OJA). While most often only one vacant position is indicated, frequently 2-5 vacant positions are indicated and several times 10 positions, up to one case each of 12 and 15 vacant positions. While not representative, this does suggest that one OJA often refers to several vacancies. Finally, a third aspect of the relation between OJA and actual vacancies concerns outdated OJA that were not taken down. When the date of the post is

Funded by
The European Union

indicated, or based on the time stamp of the child page showing the OJA, old and therefore likely outdated posts may be removed. Yet in many cases, the date is not indicated, and time stamps can be automatically reset (Plaimauer, 2024).

Questions regarding data quality may also be raised about specific information within an OJA. To begin with, there is a stark difference between the pre-defined fields and the main text that was freely entered by the user: the main text can provide much richer information but also allows for far more variety, selective omissions and outright errors than the pre-defined fields. However, selective omission likely also concerns certain pre-defined fields. For example, one can reasonably expect that missing wage information does not occur randomly (see e.g. Brenzel et al., 2014). In turn, this implies that the presence of wage information could be seen as an indicator for high demand.

6. Steps towards shortage measures

While the preceding section has discussed the raw data from an initial collection of data, they do not yet say much about shortages, even when the objective cannot be to accurately measure shortages but to gain an idea of where shortages might well be found. To this end, some (estimated) shortage measure ultimately needs to be constructed. Specifically, job vacancy rates seem like a useful measure, as they are readily understandable and are considered a good indicator (van Smoorenburg, 2024). This requires extraneous but by and large available data on employment, to which the collected data on OJA may be related. Naturally, the collected data for country will not cover all vacancies, for several reasons: some vacancies are not posted online, some platforms of OJA are inaccessible, and OJA posted on niche websites or social media are missed when focussing on the leading platforms in a specific country. In other words, the collected OJA data are only a partial reflection of vacancies and job vacancy rates based on them likely involve a substantial margin of error.

On this background, the change in job vacancy rates from, say, one quarter to another may be a less error-prone measure because certain influences and errors in the level would cancel out. For example, Vermeulen and Gutierrez Amaros (2024) highlight the problem that the share of vacancies that are posted online often grows over time. Therefore, let x_1 denote the total level of actual vacancies in quarter 1 and α_1 the share of vacancies that are posted online in quarter 1. While these components are not observed, data on OJA provide an estimate of $\alpha_1 x_1$, and $\alpha_2 x_2$ is defined anagolously. With r denoting the observed rate of growth in OJA,

$$\alpha_2 x_2 = (1 + r)\alpha_1 x_1 \qquad (1)$$

Denoting the rate of growth in the share of vacancies that are posted online by m, the relation between the two shares is $\alpha_2 = (1 + m)\alpha_1$. Using this in equation (1), α_1 cancels out and rearranging leads to

Funded by
The European Union

$$\frac{x_2}{x_1} = \frac{1+r}{1+m} \qquad (2)$$

If the growth rate m is roughly constant over time, then an estimate for $(1+r)/(1+m)$ can be derived, some information on m, possibly obtained from observing longer-term trends rather than quarter-on-quarter changes. This estimate for the growth rate of total vacancies can therefore be derived even when total vacancies and the shares of vacancies that are posted online are all unknown. This argument does not only apply at the country level but also at the level of sectors or occupations. However, the information on m would likewise have to refer to the sector or occupation in question. Needless to say, focussing on changes is only possible when there are several waves of data collection, and if the grouping of the data remains stable. The web scraping is indeed intended to continue over time, and to be progressively extended to more countries.

The need to translate and classify the raw data represents another important obstacle on the way towards useful shortage measures. As explained in Section 5, it appears more promising to focus on job titles and classify them using a standardised classification of occupations. If this system is only defined in certain languages, the first step would have to assure a high-quality translation of job titles from the various other languages into a language used for classification. The classification itself can be done using "libraries" that contain a large number of job titles or occupations and indicate which ones are synonyms. By matching a (possibly translated) job title in the collected data to some entry in a library, a synonym may be identified that appears in the classification.

While the collected data sometimes contain information on expected education levels or prior work experience, such information is mostly unavailable. One rather direct option is to try and find information on skill levels in the main text of the vacancy, searching for instances of pre-defined key words. Another, more indirect option is to rely on the classification in terms of occupations that was just outlined above. Further systems of classifications have already assigned skill levels to detailed occupations (as well as tasks performed in the occupation, according to standards such as ESCO or O*NET).

Finally, it is worth noting a potential further shortages measure based on the date when an OJA was posted. The time that elapses before a vacancy is filled (the "vacancy duration") has been used as shortage measure in practice and in academic research (e.g. Bassier et al., 2024). While the time when a vacancy is filled is not observed here, the "current age" of an OJA can be calculated: information on the date of the posting appears frequently available and, where it is missing, might often be complemented by the time stamp of the web page that only shows a single OJA. This time stamp (e.g. "last modified on") may also be found in the HTML code of websites. As mentioned before, however, both the pre-defined fields for the date the OJA was posted and the time stamps might not always be reliable, so that data quality represents a challenge also for vacancy duration as a shortage measure.

Funded by
The European Union

References

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1), S293-S340.

Antolín, P. (1994). Unemployment flows and vacancies in Spain. University of Oxford, Institute of Economics and Statistics. Discussion paper No. 158

Bassier, I., Manning, A. & Petrongolo, B. (2024). Vacancy duration and wages. LSE mimeo

Batata, A. S. (2005). International nurse recruitment and NHS vacancies: a cross-sectional analysis. *Globalization and Health*, 1, 1-10.

Beręsewicz, M., Cherniaiev, H., & Pater, R. (2021). Estimating the number of entities with vacancies using administrative and online data. arXiv preprint

Brenzel, H., Gartner, H., & Schnabel, C. (2014). Wage bargaining or wage posting? Evidence from the employers' side. *Labour Economics*, 29, 41-48.

Brown, D., Magrini, E., & Pelucchi, M. (2024). Predicting Skill Shortages with Real-Time Data: Using Online Job Adverts to Predict UK and Italian Employer Perceptions. In *Shortages of Labour and Skills* (pp. 89-98). Nomos.

Duval, M. R. A., Duval, R., Ji, Y., Li, L., Oikonomou, M., Pizzinelli, C., ... & Tavares, M. M. (2022). Labor market tightness in advanced economies. International Monetary Fund.

Evans, D., Mason, C., Chen, H., & Reeson, A. (2023). An algorithm for predicting job vacancies using online job postings in Australia. *Humanities and Social Sciences Communications*, 10(1), 1-9.

Gu, R., & Zhong, L. (2023). Effects of stay-at-home orders on skill requirements in vacancy postings. *Labour Economics*, 82, 102342.

Jackman, R., Layard, R., & Pissarides, C. (1989). On Vacancies. *Oxford Bulletin of Economics & Statistics*, 51(4).

Lambon-Quayefio, M., Asante, K., Dzansi, J. & Telli, H. (2024). Can online Job Adverts help SMEs find the Right Workers? Evidence from Ghana.

Marinescu, I., & Wolthoff, R. (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics*, 38(2), 535-568.

OECD (2022). Skills for Jobs 2022: Mapping skill requirements in occupations based on job postings data. OECD Publishing, Paris.

Plaimauer, C. (2024). OJA Analysis as Data Source for Monitoring Staffing Bottlenecks in Austria. In *Shortages of Labour and Skills* (pp. 99-108). Nomos.

Funded by
The European Union

Teo, H., Vadean, F., & Saloniki, E. C. (2022). Recruitment, retention and employment growth in the long-term care sector in England. *Frontiers in Public Health*, 10, 969098.

Tjaden, J. (2023). Web Scraping for Migration, Mobility, and Migrant Integration Studies: Introduction, Application, and Potential Use Cases. *International Migration Review* 2023, 1-18.

van Smoorenburg, M. (2024). A Comparison of Labour Market Tightness over Time and Between Countries: An Evaluation of Indicators. In *Shortages of Labour and Skills* (pp. 49-58). Nomos.

Vermeulen, W., & Amaros, F. G. (2024). How well do online job postings match national sources in European countries? OECD LEED paper 2024/02

Yeh, C., Macaluso, C., & Hershbein, B. (2022). Monopsony in the US labor market. *American Economic Review*, 112(7), 2099-2138.

Appendix

See Excel file “Database v2” for the raw data collected so far.

Funded by
The European Union

GS4S Working paper series (D7.3)

Working paper no. 3

Locating shortages in migrants' origin countries: a big data approach

About GS4S

GS4S seeks to better understand global skills shortages in selected sectors (Digital, Care and Construction) and strengthens evidence-based and multi-level policies on labour migration governance. The project provides new knowledge on alternative and equitable ways for addressing skills shortages in six regions (EU, EEA, Western Balkan, Middle East and Northern Africa, West Africa, and South/South-East Asia).

www.gs4s.eu

Funded by
The European Union