Internal and external validity of the vacancy database

Jeisson Cárdenas

ALIANZAEFI
economía formal e inclusiva

Documento de Trabajo
Alianza EFI - Colombia Científica
Abril 2020

Número de serie: WP2-2020-005

Internal and external validity of the vacancy database [1] [2]

Jeisson Cárdenas Rubio
Institute for Employment Research
University of Warwick
Coventry, United Kingdom

Abstract

This paper provides an evaluation of the internal and external consistency of the vacancy information. The consistency of the variables within the vacancy database or internal validity shows that the contradictory or inconsistent results that occurred in the Colombian vacancy database were minor, and the magnitude of these measurement errors are insufficient to bias the educational, occupational, sectorial, skills and wage analyses. The results of data representativeness or external validity were: 1) the vacancy database is not representative for a significant part of agricultural, government and armed force occupations; 2) particular caution should be taken when analysing occupations with high turnover rates as this issue might cause an overrepresentation of specific occupational groups; and, 3) self-employed individuals and informal occupations are not represented in the vacancy database. This evidence suggests that the vacancy database better represents the formal and urban Colombian labour market. Finally, the job portal information captures and expresses the Colombian economic seasons.

Key words: Labour demand, vacancy database, online job portals.

JEL classification: J23, J24, J31

[1] This working paper is part of the author's PhD thesis at the University of Warwick.
[2] E-mail: j.cardenas-rubio@warwick.ac.uk (J.Cardenas).

1. Introduction

Cárdenas (2020d) described the main characteristics of the Colombian vacancy database from 2016 to 2018. However, these results do not provide enough evidence about the validity or reliability of vacancy data for addressing unemployment and informality problems in Colombia. As is the case with data collected with other methods (e.g. surveys), the data collected from online sources have some caveats that might affect the interpretation of results. Companies can post mistaken or contradictory information; for instance, employers might request an engineering professional with just a high school diploma or a full-time engineering professional with an extremely low salary. Moreover, errors in posted information might arise from the data mining processes. The algorithms created in Cárdenas (2020c) might fail. For instance, the algorithm that looks for patterns in job descriptions might confuse some words, and incorrectly create variables of a university degree or job experience, among others. Consequently, errors or biases might arise in the information and affect the internal and external consistency of the vacancy database, hence the labour market analysis might arrive at wrong conclusions. Thus, this paper tests the internal and external validity of the vacancy information.

Internal validity refers to the consistency of the variables within the vacancy database (Henson, 2001; Streiner, 2003). In ideal conditions, the results from a variable in the vacancy database should not contradict the findings from other variables in the same database; otherwise, the results will be unreliable. One straightforward way to address this issue is to compare the results of different but related variables. Therefore, the second section of this paper tests the internal validity of the vacancy database via cross-tabulations and wage distribution analysis.

Internal validity is a crucial aspect to consider before drawing any conclusions about labour demand from the vacancy database. Moreover, to establish result consistency from the vacancy database within a particular economic context (external validity) is another relevant factor to consider before drawing any conclusions about Colombia's labour market (Kureková et al. 2014). External validity, specifically, refers to possible biases or representativeness issues in the data (Rasmussen, 2008; Stopher, 2012).

Logically, all sources of information have limitations. For instance (as mentioned in Cárdenas, 2020a), in Colombia the current sectoral surveys carried out by DANE do not provide detailed

information about human capital, such as occupational structure or the skills required in each position. Web-based information might help to fill this gap. However, the online sources utilised for the database in this study also have limitations.

Given the nature of these online sources, job vacancy information might describe a particular segment of the labour market. The external validity of results depends on which kinds of vacancies are being published online for the country of interest. To test external validity, it is necessary to process and compare the results from other sources of information (e.g. household surveys) with the vacancy database results. Therefore, Section 3 discusses the representativeness of the Colombian vacancy database by

a. Categorising the household labour survey (GEIH) according to ISCO-08 categories.
b. Comparing the Colombian vacancy data set with official national labour statistics.

Finally, Sections 2 and 3 propose a framework to evaluate the representativeness of the vacancy database for each occupation at different levels of disaggregation (e.g. four-digit ISCO level):

- When testing the internal consistency of the information for a specific occupation are there minors or null errors?
- If yes, are the distribution of wages in the vacancy database for that particular occupation similar to the distribution of wages in the household survey?
- Can similar seasonal trends be observed in the level of employment in the household survey and in the level of job vacancies?
- Can opposite seasonal trends be seen in both the level of unemployment in the household survey and the level of job vacancies?
- Do lagged effects exist between the number of job advertisements and new hires?

This framework is particularly useful for countries such as Colombia, where testing and comparing the representativeness of a vacancy database built from online sources is more challenging because labour demand information collected by traditional methods (such as vacancy surveys) is scarce.

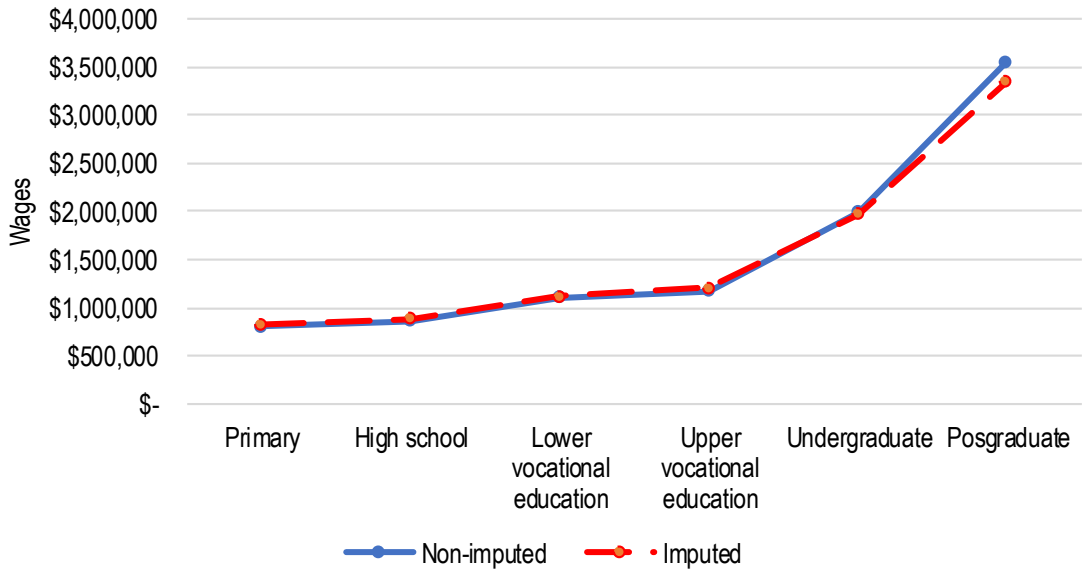
2. Internal validity

Establishing the internal validity or the internal consistency of a database implies that the results from a variable should not contradict the findings from another variable (Henson, 2001; Streiner, 2003). If employers demand engineers or economists, for instance, most of the vacancies for those job positions should also demand people with at least some university educational level (when the educational level is mentioned in the job advertisements). Additionally, according to human capital theory higher salaries should be positively correlated with a higher level of human capital (see Cárdenas, 2020a); otherwise, the results would be contradictory. In this case, job portals might not be a reliable source of labour demand (skill mismatch) information, or the algorithms developed in Cárdenas (2020c) might be failing. To test the internal validity of the vacancy database involves the comparison of different but correlated variables.

2.1. Wage distribution by groups

One straightforward way to prove the internal consistency of the vacancy database is comparing the average salary of different population groups. Usually, vacancies that require a person with a high level of education should pay higher wages than vacancies that ask for a person with a relatively low level of education (see Cárdenas, 2020a). Figure 1 shows the average imputed and non-imputed salaries by educational level. As expected, vacancies that require people with a low level of education pay lower wages than vacancies that ask for people with a high level of education. On average, jobs that require a basic level of education (primary or high school) pay a salary of 829,000 pesos monthly (around £207), while jobs for undergraduates and postgraduate pay 1,975,040 pesos and 3,350,764 pesos (around £494 and £838), respectively. Moreover, as mentioned in Cárdenas (2020d), the differences between imputed and non-imputed wages are minimal. This comparison suggests two facts: 1) the imputation process carried out in Cárdenas (2020c) does not significantly affect the wage distribution variable. Hence, imputed wages (the whole database) can, potentially, be used for the analysis of labour demand; and, 2) the vacancy information contains consistent results at least for the education and wage variables.

Figure 1. Education and wages (pesos)[3]

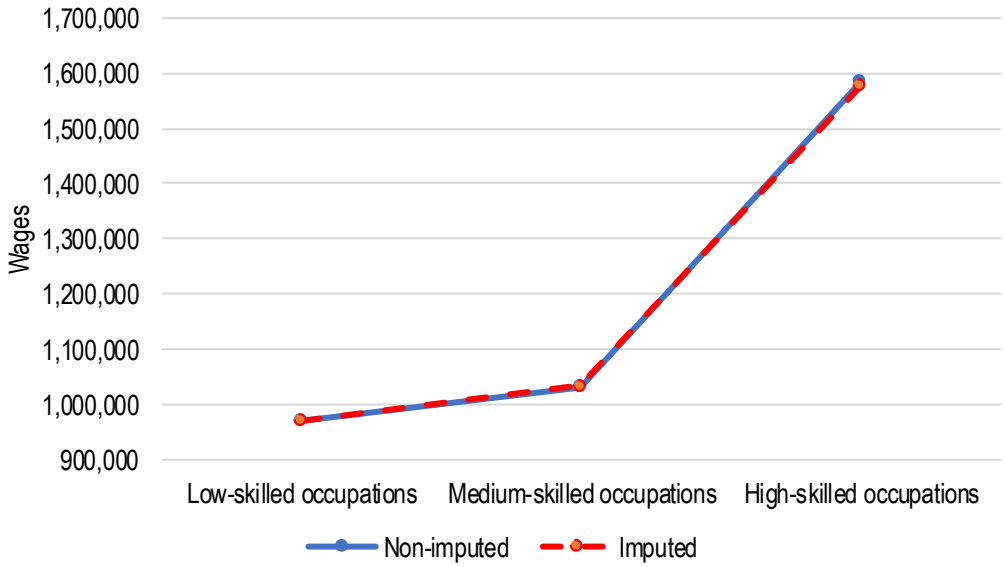


Source: Vacancy information 2016 - 2018. Own calculations.

Similar to the education variable, it is logical to expect that high-skilled jobs tend to pay higher salaries than low-skilled jobs. Figure 2 presents the average wages (imputed and non-imputed) that employers are willing to pay for high, medium and low-skilled occupations. On average, the wage for a low, medium and high-skilled occupation is around 970,000 (around £242), 1,034,000 (around £258) and 1,577,000 pesos (around £394), respectively. Moreover, the imputed and non-imputed wage variables overlap for each occupational group. Thus, there is a positive relationship between wages and the degree of complexity of an occupation.

[3] Given the relatively low frequency of Specialisation, Master and PhD degrees, these categories were grouped into a single category named "Postgraduate".

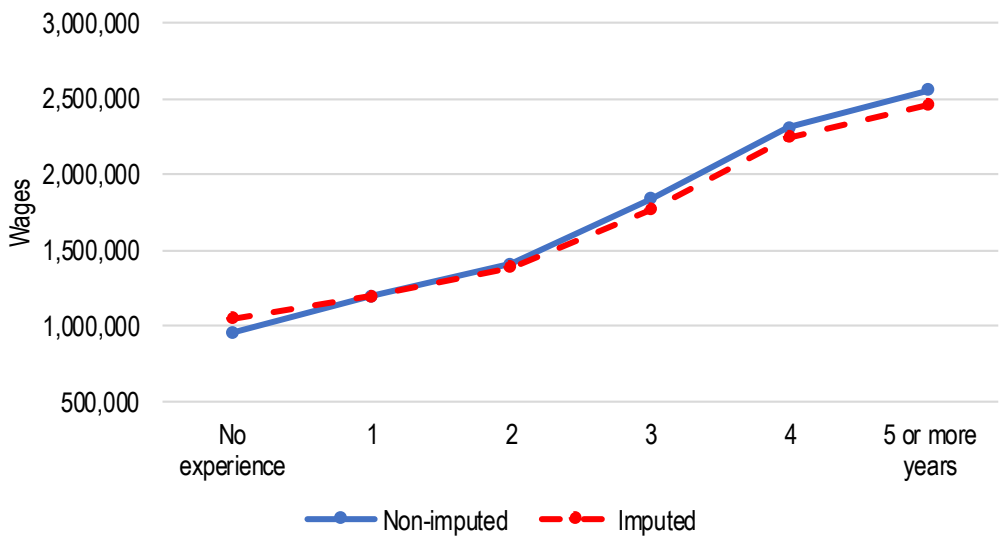
Figure 2. Occupations and wages (pesos)



Source: Vacancy information 2016 - 2018. Own calculations.

To provide more evidence regarding the consistency of human capital requirements and average wages in the vacancy database, Figure 3 presents the average wage (imputed and non-imputed) by experience requirements. Vacancies that do not require labour experience pay, on average, a salary of 1,045,000 pesos monthly (around £261), while vacancies that require five or more years of experience pay, on average, 2,457,000 pesos (around £838) monthly.

Figure 3. Years of experience and wages



Source: Vacancy information 2016 - 2018. Own calculations.

Therefore, the above evidence suggests that the information regarding human capital demand in Colombia is consistent with wage information. Consequently, this also means that the web scraping, text mining, imputation and classification processes used to collect the vacancy data provide consistent results with which to analyse the labour market. However, the wage variable is an insufficient indicator to test internal data consistency because the vacancy distribution between groups is also required to determine the internal consistency of the vacancy database.

2.2. Vacancy distribution by group

Not unlike the wage variable, the distribution of vacancies should provide consistent results. As previously mentioned, if employers demand engineers, economists, or any other occupation that implicitly requires an undergraduate diploma, then most of the vacancies for such job positions should also demand people with at least some university educational level. In addition, the skills listed in Cárdenas (2020d) should correspond to their related occupations; SQL programming skills, for instance, should correspond to programmers and other related occupations, while it is unlikely that programming skills would correspond to chefs, taxi drivers, and plumbers.

Table 1 reveals job distribution according to educational requirements, for occupations following the OECD's (2017c) categories. On the one hand, according to Column 1, around 41.1% of the jobs that require a basic education level (primary school) correspond to low-skilled occupations,

while 56.3% and 2.7% correspond to medium-skilled and high-skilled occupations, respectively. On the other hand, only 1.6% of the jobs that require a postgraduate diploma correspond to low-skilled occupations, while 5.4% and 93.0% correspond to medium-skilled and high-skilled occupations, respectively. This result suggests that the information regarding human capital requirements in the vacancy database is consistent. Indeed, in Table 1 the red zones indicate the lowest cell values, as the level of education increases the percentage of low and medium-skilled occupations decreases. The green zones indicate the highest cell values, and as the level of education increases the percentage of high-skilled occupations increases.

Table 1. Occupational structure by education

Occupation	Primary	High school	Low vocational education	Higher vocational education	Undergraduate	Postgraduate
Low-skilled	41.1%	29.1%	15.0%	11.4%	6.2%	1.6%
Medium-skilled	56.3%	42.1%	32.8%	27.4%	14.0%	5.4%
High-skilled	2.7%	28.8%	52.2%	61.1%	79.8%	93.0%
Total	100%	100%	100%	100%	100%	100%

Source: Vacancy information 2016 - 2018. Own calculations.

Despite the variable sector containing a significant number of missing observations (see Cárdenas, 2020d), this variable might provide more evidence regarding the constancy of the vacancy database. A sector might demand different occupations that are not directly related to the main activity of the industry. For instance, the finance sector hires finance managers and finance analysts, among other associated occupations; however, this sector might also demand security guards and sales representatives. Notwithstanding the wide range of occupations required by each industry, differentiating patterns should exist between the occupational structure of one sector of labour demand and another. The vacancy data should show, for instance, that the finance sector demands relatively more finance analysts than the agriculture sector; otherwise the vacancy information might contain significant errors that can prevent a researcher from drawing academic or public policy recommendations.

Given the number of groups of occupations by industry, Table 2 shows some of the most notable cases of the labour demand occupational structure (at a four-digit level) by sector (at a one digit-level). For instance, Column 1 from Table 2 presents the ten occupations most demanded by companies related to "Real estate activities". As can be seen, the second most required occupation for this category is "Real estate agents and property managers", while in the other sectors this occupation is not frequently demanded. Companies related to "Accommodation and food service activities" frequently demand "Kitchen helpers", "Cleaners and helpers in offices, hotels and other establishments", and "Stock clerks". From Table 2, it can be concluded that occupations and sector variables have an expected correlation which suggests that the occupational and industry variables, in general, provide consistent results.

Table 2. Top 10 occupational labour skills in demand by sector

#	Real estate activities	Accommodation and food service activities	Wholesale and retail trade; repair of motor vehicles and motorcycles	Manufacturing	Transportation and storage
1	Commercial sales representatives	Kitchen helpers	Commercial sales representatives	Commercial sales representatives	Stock clerks
2	Real estate agents and property managers	Cleaners and helpers in offices, hotels and other establishments	Sales demonstrators	Sewing machine operators	Mail carriers and sorting clerks
3	Accountants	Stock clerks	Stock clerks	Cashiers and ticket clerks	Commercial sales representatives
4	Administrative and executive secretaries	Commercial sales representatives	Telephone switchboard operators	Stock clerks	Freight handlers
5	Telephone switchboard operators	Waiters	Security guards	Accountants	Building construction labourers
6	Building architects	Cooks	Cashiers and ticket clerks	Security guards	Security guards
7	Sales and marketing managers	General office clerks	Shop sales assistants	Shop sales assistants	Accountants

8	Stock clerks	Receptionists (general)	Waiters	Production clerks	Messengers, package deliverers and luggage porters
9	Receptionists (general)	Cashiers and ticket clerks	Crane, hoist and related plant operators	Services managers not classified elsewhere	Car, taxi and van drivers
10	Survey and market research interviewers	Chefs	Accountants	Mail carriers and sorting clerks	Administrative and executive secretaries

Source: Vacancy information 2016 - 2018. Own calculations.

The skill information is one of the potential advantages of the vacancy database (see Cárdenas, 2020c, 2020d). Thus, it is essential to test the internal consistency of the "skills" variable. Testing this variable might be challenging because some skills (generic skills) are demanded regardless of the level of education, wage or occupation. Additionally, it might take a considerable time to test the consistency of each skill[4]. To avoid these issues, ten skills explicitly related to an occupational group were chosen to test the internal validity of the "skills" variable. For instance, most of the jobs that require SQL or JavaScript programming skills should correspond to programmers and related occupations.

Table 3 shows the occupations with the highest demand for ten ESCO skills. For instance, the occupations with the highest demand for SQL programming skills are "Web and multimedia developers", followed by "Systems analysts and database designers and administrators". Similar occupations are demanded when employers require JavaScript skills. In contrast, when "Carpentry skills" are needed, the most frequently requested occupation is "Carpenters and joiners", followed by "Odd job persons", and "Mechanical engineering technicians". Additionally, "Generalist medical practitioners", "Nursing professionals" and "Specialist medical practitioners" are the most frequently demanded occupations when employers require epidemiology skills. This evidence suggests that skill information is consistent with occupation variable which, in turn, provides corresponding results with the educational and wage variables.

[4] Around 4,000 thousand were identified in the vacancy descriptions, see Cárdenas (2020d).

Table 3. Top 10 occupational skill categories

#	SQL	JavaScript	Carpentry	Epidemiology	Mechanics
1	Web and multimedia developers	Web and multimedia developers	Carpenters and joiners	Generalist medical practitioners	Mechanical engineering technicians
2	Systems analysts	Systems analysts	Odd job persons	Nursing professionals	Electrical mechanics and fitters
3	Database designers and administrators	Engineering professionals not classified elsewhere	Mechanical engineering technicians	Specialist medical practitioners	Mining engineers, metallurgists and related professionals
4	Information and communications technology user support technicians	Information and communications technology user support technicians	Stock clerks	Physiotherapists	Crane, hoist and related plant operators
5	Engineering professionals not classified elsewhere	Web technicians	Production clerks	Dentists	Mechanical engineers
6	Software developers	Software developers	Commercial sales representatives	Biologists, botanists, zoologists and related professionals	Motor vehicle mechanics and repairers
7	Electronics engineers	Graphic and multimedia designers	Building construction labourers	Health professionals not classified elsewhere	Production clerks
8	Information and communications technology operations technicians	Electronics engineers	Sewing, embroidery and related workers	Office supervisors	Mail carriers and sorting clerks
9	Web technicians	Building architects	Assemblers not classified elsewhere	Other artistic and cultural associate professionals	Heavy truck and lorry drivers
10	Electronics engineering technicians	Telecommunications engineering technicians	Information and communications technology installers and servicers	Chemists	Welders and flamecutters

Source: Vacancy information 2016 - 2018. Own calculations.

All the evidence presented above suggests that the vacancy database is internally consistent. However, it is important to note that every database regardless of its sources might have some errors[5]. For instance, Table 1. Occupational structure by education

shows that around 2.7% (3,685 out of 136,479 observations) of the jobs that required education at primary school level correspond to high-skilled occupations. This result is suspicious because usually high-skilled jobs require a higher educational level. Indeed, a closer look in the vacancy database shows that a portion of these 3,685 jobs was misclassified[6].

However, many mistakes are easy to identify and correct. In fact, one of the most critical advantages of scraping the data directly from job portals is that the researcher has the possibility to evaluate and correct possible mistakes in the gathered information. Algorithms might fail and might provide contradictory or inconsistent results, however, the quality of the data created (i.e. dummy variables such as education and experience, among others) can be tested against the original data (i.e. job description, job title, etc.), and the algorithms can be easily refined until they provide a certain level of consistent results. For the Colombian vacancy database, the evidence shows that contradictory or inconsistent results are minor, and the magnitude of these measurement errors are not large enough to bias educational, occupational, sectorial, skills and wage analysis. Thus, the above comparison of variables provides more confidence in the results derived from the vacancy database.

3. External validity

Section 2 illustrates that the vacancy database provides consistent internal outcomes. This result is important because it proves that non-traditional sources such as the Internet and methods for collecting and organising (such as web scraping and text mining) might be used to analyse

[5] In household surveys when people are asked about their wages, they can provide wrong information, or the interviewer might write an incorrect value. However, the depuration processes carried out by the Bureau of Statistics guarantees that these measurement errors are minor and do not bias household survey results at a certain disaggregation level.

[6] For instance, some of these jobs demanded primary school teachers and the text mining algorithm misunderstood because the pattern "primary school" was in the job description, hence the educational requirement was wrongly assigned to "primary".

labour demand. Nevertheless, internal validity does not entirely prove the limits of the vacancy database. A database can provide consistent internal results, yet the data might not properly represent a population group (sample error), hence academic or public policy conclusions drawn from that data might be biased.

Thus, the external validity or representativeness of a database is one of the essential elements to consider before drawing any conclusions based on that particular database (Stopher, 2012; Rasmussen, 2008). Despite the importance of testing the external validity of vacancy information, different authors have derived conclusions from job portal information without a careful analysis regarding data representativeness. For instance, Kennan et al. (2008) investigated what knowledge, skills, competencies and personal characteristics employers list in online job advertisements for Information System graduates. While Backhaus (2004), asked: How is information in corporate descriptions (job advertisements) distributed across the various information categories?; and, How do firms attract the curiosity of desired potential workers by selecting the appropriate words in online job advertisements? In addition, Kureková et al. (2016) studied what combination of cognitive and non-cognitive skills were required by employers for low- and medium-skilled occupations in three European countries: the Czech Republic, Denmark and Ireland.

The three articles cited above used vacancy databases and possessed a fundamental assumption; which is, that online job advertisements are a valid representation of labour demand. They assume that information from the Internet is representative of the whole economy or a specific sector. However, since this information does not come from a sampling frame, these sources may not be representative given the penetration of internet usage (Štefánik, 2012). According to Carnevale et al. (2014), the main source of bias in a job vacancy database might be due to differences in Internet access among job applicants in terms of education level or skills.

Thus, more information does not guarantee better results. In consequence, not knowing the direction of the bias might provide the wrong conclusions or limit the scope of the studies. A possible bias in the gathered information can affect vacancy analysis in the following two ways: 1) job portals could publish only high-skilled jobs, while printed or voice-voice vacancies might correspond to middle- or low-skilled jobs. This possible source of bias might (in this case)

overestimate the labour demand for high-skilled jobs, and educational providers might saturate the labour market with more high-skilled people than the labour demand requires. In this thesis, this bias is named the "selection bias"; and/or, 2) the vacancies posted in job portals might not properly describe the characteristics (e.g. the skills) required by employers. Jobs portals could tend to publish particular information to attract the attention of those that use the Internet, while printed or voice-voice vacancies might publish different information (such as skills or educational requirements) to attract those people that use this medium to search for jobs. In this thesis, this bias is named the "description bias".

Concerns regarding "description bias" were in part answered in the previous section. As observed, job requirements such as skills, education, etc., correspond to the expected requirements for each occupation. "Description bias" seems implausible in the vacancy database because occupational requirements do not depend on the way the vacancy is advertised. For instance, the skills needed for a plumber do not change because the vacancy was posted online or transmitted voice-voice—the general tasks of a plumber are the same[7].

However, the vacancy data per se cannot answer when it is appropriate to provide more or less of a particular skill in response to labour demand. To accurately address this issue, it is necessary to identify any possible "selection bias". Job portals might advertise more or fewer vacancies for a specific occupation regardless of the economic season or trends[8].

Testing the "selection bias" might be challenging. As mentioned in Cárdenas (2020a), official labour demand surveys are characterised by a sampling frame (based on a census of people, companies, etc.) which ensures the data and results are representative of a certain population. Consequently, given this statistical design, it is relatively easy to calculate the degree of representativeness in official household and sectorial surveys. Nevertheless, to test vacancy

[7] Subsection 3.1.2 provides more evidence of this point, suggesting that "description bias" is not a predominant issue in the vacancy database. Thus, the vacancy data can provide valuable answers about what people should be trained in at a low cost (time and money).

[8] For instance, employers might opt to use job portals to collect CVs and store them in their databases (see Cárdenas, 2020a) regardless of whether it is a period where more people are hired or not. Consequently, vacancy information from job portals might not be a useful source to determine trends, seasonal or cyclical changes in labour demand.

data representativeness is not an easy task given that this information is not collected based on a sampling frame. Ideally, to examine the data representativeness of information from job portals an updated census of vacancies which details the characteristics of human resource requirements is required. Nevertheless, to carry out this census is costly. Thus, countries such as Colombia (with a restricted budget, see Cárdenas, 2020a) do not have a census of vacancies or any similar labour demand information to refer to. This absence of a vacancy census or survey makes it difficult to know the limits of job portal information.

One way to address this issue is by comparing vacancy information with household surveys. Indeed, Štefánik (2012) compares the most popular job search website vacancies for tertiary education graduates in Slovakia with a labour force survey for the same educational group. As Štefánik (2012) points out, this approach assumes that occupational and sector structures in the vacancy database are similar to employment distribution by occupational and sector groups. According, to this method, a vacancy database adequately represents labour demand information if there is a sufficiently high correlation with employment surveys. In aggregated terms, comparing vacancy data with household surveys can provide relevant insights regarding the representativeness of job portal information. For instance, by comparing the number of vacancies with the level of employment over time it is possible to determine if job portal data adequately captures the behaviour of companies during economic cycles and seasons. It is expected, for instance, that the level of vacancies sharply increases at the end of each year given the increase in economic activities during that period, or decreasing the number of vacancies during periods of economic recessions and vice versa.

Moreover, the comparison between aggregated (one or two-digit level) occupational structures of the vacancy database with occupational groups from household surveys might identify a possible under/over-representation of specific occupational groups in the vacancy database. At a one or two occupation digit level (the household at a more disaggregated level such as a four-digit ISCO might have representativeness problems), both the vacancy and the household data should have a similar occupational distribution if job portal information adequately covers all occupational groups in the economy. Otherwise, vacancy information might over/under-represent a particular occupational group.

One alternative explanation for the difference between the occupational structure of vacancy and household surveys might be that the labour market has a relatively high skill shortages problem. Given the existence of mismatches in the labour market, labour demand information might not coincide with labour supply information. This argument might justify why detailed comparisons between vacancy and household data are an improper method to test vacancy data representativeness. However, in aggregated terms (one or two occupational digit level) the differences between the labour structure of labour demand and supply might not be properly explained by the hypothesis of skill mismatches. For instance, a higher participation of "Professionals" (one-digit level ISCO, major group) in the vacancy database compared with the information from household surveys would suggest, under the hypothesis of skill mismatches, that the country has a significant shortage of any professionals. Nevertheless, this explanation does not seem plausible because if there were such evident skill shortages the wages of professionals would be significantly higher, and the unemployment rate would be considerably lower than other occupational groups. With such obvious evidence concerning labour market mismatch, education and training providers, the government and, in general, people should react to this imbalance and correct the issue. For these reasons, the mismatch hypothesis might not explain occupational differences at a one or two-digit level between vacancy and household survey information.

Thus, to compare the vacancy database at an aggregated level (i.e. major occupational groups) with the information from household surveys is the most straightforward approach to identify possible biases in job portal information. However, to conduct a more detailed comparison to test the data representativeness of vacancy information between the vacancy database and a household survey might be problematic. In concordance with Kureková et al. (2014), household surveys provide information regarding labour supply which is composed of the number of job matches (level of employment, see Cárdenas, 2020a) and the number of people unemployed, while job portal information is the total of the net, and replacement, labour demand.

Therefore, a direct comparison with household surveys at a detailed level (i.e. ISCO minor groups) might not be a suitable proxy to test the data representativeness of the vacancy database. Besides, vacancy information might contain and reflect seasonal or future changes that might not match the current labour supply (the possibility of skill mismatches occurring). For instance, as mentioned in Cárdenas (2020a), the rapid emergence of modern devices (e.g.

computers, smartphones, etc.) have introduced new technologies in the labour market to perform different jobs, such as programmers, data analysts, among others. These accelerated changes have been reflected in the labour demand for skills, and have been documented by different authors such as Acemoglu and Autor (2011). However, the current employment structure might require more time to reflect those changes due to (for instance) the time that people need to be trained and offer specific skills.

Considering the advantages and the limitations of comparing the vacancy database with household surveys, the following subsection evaluates Colombian vacancy data representativeness by comparing the vacancy database with Colombian household surveys.

3.1. Data representativeness: vacancy versus household survey information

As mentioned above, the most straightforward way to evaluate vacancy data representativeness is by comparing the result of the occupational structure or employment trends of this source of information with the results from household surveys. The Statistics Office of Colombia (DANE) has carried out a monthly cross-sectional household survey named "*Gran Encuesta Integrada de Hogares*" (GEIH) since 2006 (see Cárdenas, 2020a)[9]. The GEIH is the main source of official labour market information in Colombia.

3.1.1. Occupational structure

At the time this thesis was written, the DANE classified people's occupations by the SOC 1970[10]. Perhaps one of the reasons the DANE has not updated their labour supply statistics with ISCO-08 is because the Colombian statistics department still uses manual codifiers (a group of people) to code job titles one by one in its household surveys. As explained in Cárdenas (2020c), the manual classification of job titles is a time-consuming task; consequently, to update all of the

[9] With a total sample size of approximately 23,000 households monthly, this source of information measures the characteristics of the Colombian workforce. The GEIH collects monthly data representative at national, rural and urban levels, quarterly data representative at a cities level: Bogotá, Medellín AM, Cali AM, Barranquilla AM, Bucaramanga AM, Manizales AM, Pasto, Pereira, Cúcuta, Ibagué, Montería, Cartagena, Villavicencio, Tunja, Florencia, Popayán, Valledupar, Quibdó, Neiva, Riohacha, Santa Marta, Armenia and Sincelejo.

[10] This classification was created in 1970 by the Minister of Labour and Social Protection and SENA (Cabrera et al. 1997).

household historical records according to ISCO-08 via manual codifiers would require a considerable amount of time and money.

Both the manual classification and the use of outdated (and sometimes not well-defined) classifications might be a source of measurement errors. Manual coders might differ to the official criteria to classify a job title. Moreover, an outdated classification might not well-distinguish some occupational groups. For instance, the SOC 1970 has the two following categories at a two-digit level: code 53 (cooks, waiters, bartenders and waiters), and code 77 (food preparation workers: bakers, slaughterers, butchers, etc). Consequently, manual coders might not know how to classify a job title such as "Chef" or "Kitchen assistants". The codification might depend on the criteria of each manual coder. In fact, there are codification problems in the GEIH; workers with the same job title (such as "Fried food cook") have different occupational codes (either 53 or 77).

Cárdenas (2020c) shows that despite the relatively large amount of job titles the Colombian vacancy database is classified automatically using ISCO-08, which is (at this moment) the most up-to-date occupational classification provided by the ILO. Given the advantages of upgrading the current labour supply classifications, the following subsections outline how job titles in the GEIH can be automatically classified according to ISCO-08 to compare supply and demand occupational structures.

3.1.1.1. Categorising GEIH according to ISCO-08 categories

The GEIH requests the job title for each formal or informal worker. Moreover, all unemployed people are asked about the job position that they are looking for, and unemployed people, that have worked in the past, are asked about their last job position. Consequently, with questions about job titles and the codification of those job titles it is possible to gather information about the occupations for three different groups:

1) Individuals working in formal employment;
2) Unemployed individuals where occupation refers to the occupation they seek to work in;
3) Individuals working in informal employment.

The procedures described in Cárdenas (2020c), Section 4, were carried out to classify the job titles of the GEIH. Briefly, around 320,000 unique job titles received an occupational code (ISCO-08) by implementing a manual codification, CASCOT and a machine learning algorithm (as described in the previous Cárdenas, 2020d). Once the labour supply information was coded according to ISCO-08, it was possible to carry out the comparison between labour demand and supply information. In total, 419 occupational groups (at a four-digit level) were found in the GEIH.

3.1.1.2. Comparing supply and demand occupational structures

Figure 4 shows the percentages of potential job placements (hereafter "job placements" or "job vacancies") from the vacancy database, and the employment level in Colombia from the GEIH—all figures are arranged according to occupational group at a four-digit ISCO level. Superficially, the chart suggests that a certain level of correlation exists between labour demand and labour supply information. Indeed, the Pearson correlation coefficient is 0.34. Yet, a more detailed comparison between the labour demand and supply information reveals three facts: 1) some occupations do not appear in the vacancy data, but are found in the GEIH data; 2) conversely, no occupations are listed in the vacancy data that do not appear in the labour supply database; and, 3) despite the positive correlation between the supply and demand occupational structures, the vacancy database tends to possess a relatively higher share of technicians and associate professionals and clerical support workers (ISCO major groups 4 and 5), while the GEIH tends to possess a relatively higher of share of skilled agricultural, forestry workers and fishery workers, craft and related workers, plant and machine operators, and assemblers and elementary occupations (ISCO major groups 6, 7, 8 and 9).

First, the vacancy database does not contain information about every occupational group in the Colombian economy. Most of the occupations that are not listed in the vacancy database correspond to the military (such as commissioned and non-commissioned armed forces officers, other ranks), agriculture (animal producers, mixed crop and animal producers, inland and coastal waters fishery workers), or political and social leaders (social welfare managers, senior government officials). This is understandable, given the online sources of vacancy information and Internet penetration rates in certain zones or sectors of the country (e.g. rural zones). Thus,

the vacancy database is not representative for—at least—a significant part of agricultural, government and armed force occupations.

Second, the fact that no occupations are listed in the vacancy data that do not also appear in the labour supply database shows that information from the Internet corresponds, or does not differ from, official national labour market information. For instance, vacancies are not found for nuclear engineers and astronauts, among other occupations, because in Colombia these occupations do not have a market, so there should not be vacancies for these kinds of jobs[11]. This result suggests that online sources of information do not have a surplus of "unreal" or "inappropriate" labour demand information according to the Colombian context.

Third, the vacancy database has a significantly higher share of commercial sales representatives, telephone switchboard operators, stock clerks and sales and marketing managers, compared to the GEIH household survey. The high turnover rate of these occupations might explain this issue. Indeed, well-known business platforms such as LinkedIn detail that occupations related to marketing, research, media and communications, support and human resources are amongst those with the highest turnover rates (Linkedin, 2019).

Despite the possibility of higher turnover rates, labour demand (vacancy) and labour supply (household information) display similar patterns. For instance, commercial sales representatives (ISCO code 3322) account for around 15% of job placements. A similar peak (but of lesser magnitude) is observable in the labour supply information. The same pattern applies for accountants (2411), shop sales assistants (5223), sales and marketing managers (1221), mail carriers and sorting clerks (4412), among others. Consequently, the high job placement share of these occupations is not only due to high turnover rates, these roles also represent a relatively high portion of Colombian workers. Therefore, the peaks in job placement distribution do not provide strong evidence against data representativeness. On the contrary, this evidence

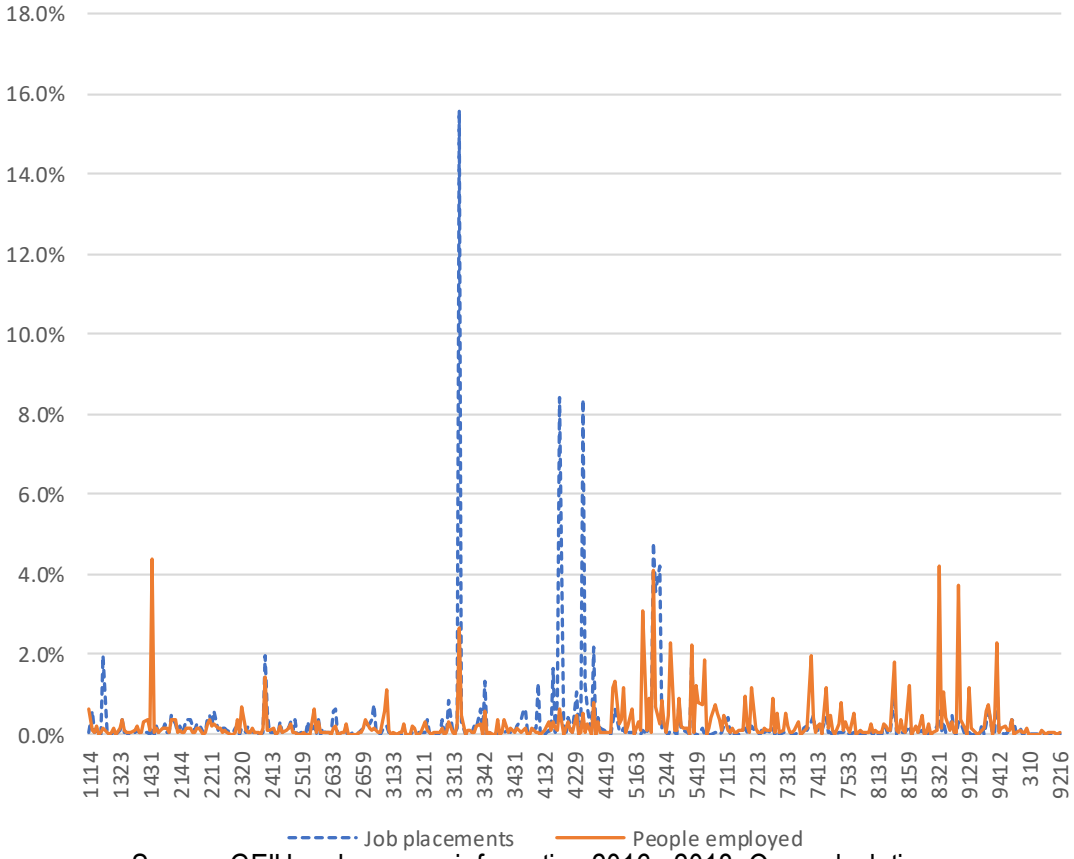
[11] Unless an industry arises that starts demanding such occupations, in which case there would be no individuals capable of carrying out the tasks required for these new occupations. However, it is not common to observe this phenomenon and this last argument is less plausible given the relatively short period of the data collected for this thesis.

suggests that the vacancy data is correlated with labour supply information and some occupations might experience overrepresentation due to higher turnover rates.

In contrast, it is unsurprising that the GEIH tends to have a relatively higher share of skilled agricultural, forestry workers and fishery workers, craft and related workers. As mentioned above, the low Internet access in certain zones or sectors of the country might negatively affect the number of jobs advertised on job portals. Moreover, the GEIH shows a relatively and considerably higher concentration of retail and wholesale trade managers (1420) and services managers not classified elsewhere (1439). A closer look at the job titles demonstrates that these occupations correspond to self-employed people that open and administrate their own businesses (for instance, a mini-market, a cafeteria, etc.). Consequently, self-employed and "business owner" occupations do not tend to be frequently announced through job portals. In fact, in Colombia these occupations tend to be found in the informal economy (see Cárdenas, 2020a). By only considering formal workers in the GEIH, the share of retail and wholesale trade managers (1420) falls to 0.4% and the Pearson correlation coefficient between labour demand and labour supply information increases to 0.39.

The above comparison between labour demand and labour supply information demonstrates at least three facts: 1) the vacancy database is unrepresentative for a significant proportion of agricultural, government and armed force occupations; 2) despite the high turnover rates of some occupations, labour demand and labour supply demonstrate similar patterns. However, special caution should be taken when analysing occupations with high turnover rates. This issue might cause an overrepresentation of certain occupational groups; and, 3) self-employed and "business owner" informal occupations are not represented in the vacancy database.

Figure 4. Job placements and employment distributions by occupational group (ISCO-08)



Source: GEIH and vacancy information 2016 - 2018. Own calculations.

3.1.2. Wage distribution of labour demand and supply information

The distribution of wages can be used as an indicator to test the representativeness of the vacancy database. It can be expected that the shape of wage distribution in the vacancy database should be similar to the distribution of wages of the labour supply. It is not expected that both the vacancy and the GEIH wages display the same distribution because the vacancy database contains information regarding labour demand and the GEIH survey collected information regarding the supply. Consequently, several reasons might explain the differences between the vacancy database's and the GEIH's wage distributions.

For instance, the vacancy database contains the initial wages that employers are willing to pay for a particular occupation, while the household survey contains a final salary figure which is agreed after a negotiation process between workers and employers. Given this bargaining process, the distribution of wages in the vacancy database for an occupation might be lower

than salaries contained in the GEIH. In contrast, skill shortages might explain why the distribution of wages in the vacancy database for an occupation might be higher than wages in the GEIH. However, it is not expected that the bargaining process, skill mismatches, etc., create significant differences between the shape of the distribution in the vacancy and GEIH datasets.

It would be difficult to explain, for example, that for a given occupation the wages in the vacancy database are negatively skewed, while the wages in the GEIH are positively skewed. One possible answer, in this case, is that the labour market is affected by relatively high skill shortage problems, and that, given these mismatches, wage distributions might not display a similar shape. However, and as mentioned above, this argument is not enough to explain the observed differences because if there are such evident and notorious skill shortages then educational and training providers, the government and, in general, people would have reacted to correct the issue. Consequently, what potentially might explain considerable differences in the shape of the distribution of wages are significant biases in the vacancy database.

Figure 5 compares the imputed and non-imputed wage distribution of vacancies (the long-dashed and dash-dotted lines, respectively), and the wage distribution of total and formal workers in the GEIH (solid and dashed line, respectively)[12]. The comparison of the distribution of wages reveals four facts. First, regardless of the source of the information, high-skilled occupations tend to pay higher salaries than low-skilled occupations. For instance, the median of the wages in the vacancy and the GEIH database for managers are 1,250,000 (non-imputed) pesos (around £312) per month, 1,614,371 (imputed) pesos (£403), 1,326,000 (total workers) pesos (£331) and 1,500,000 (formal workers) pesos (£375). In contrast, the median of the imputed and non-imputed wages in the vacancy and the GEIH database (total and formal workers) for elementary occupations is 737,700 pesos (£184) per month. This evidence confirms what is mentioned in the previous section, information regarding the human capital demand in Colombia is consistent with wages information.

Second, workers' salaries (GEIH) and job placement wages display a similar shape. Indeed, in most cases, wage distributions almost overlap. This comparison between wage distributions

[12] Given the large number of occupational groups and the representativeness issues of the GEIH at four digit-level the graphs are presented at one-digit ISCO.

demonstrates that salaries posted in job portals share a similar distribution with wages reported by Colombian workers in the official labour supply survey (GEIH). Moreover, the wage distributions of formal workers are more akin to the distribution of vacancy wages; for instance, the salary distribution for craft and related trades for the total number of workers is further to the left than for formal workers, and for vacancy (job placement) wage distributions. Consequently, the wages of informal workers are significantly lower than the vacancies and formal workers' wages (see Cárdenas, 2020a).

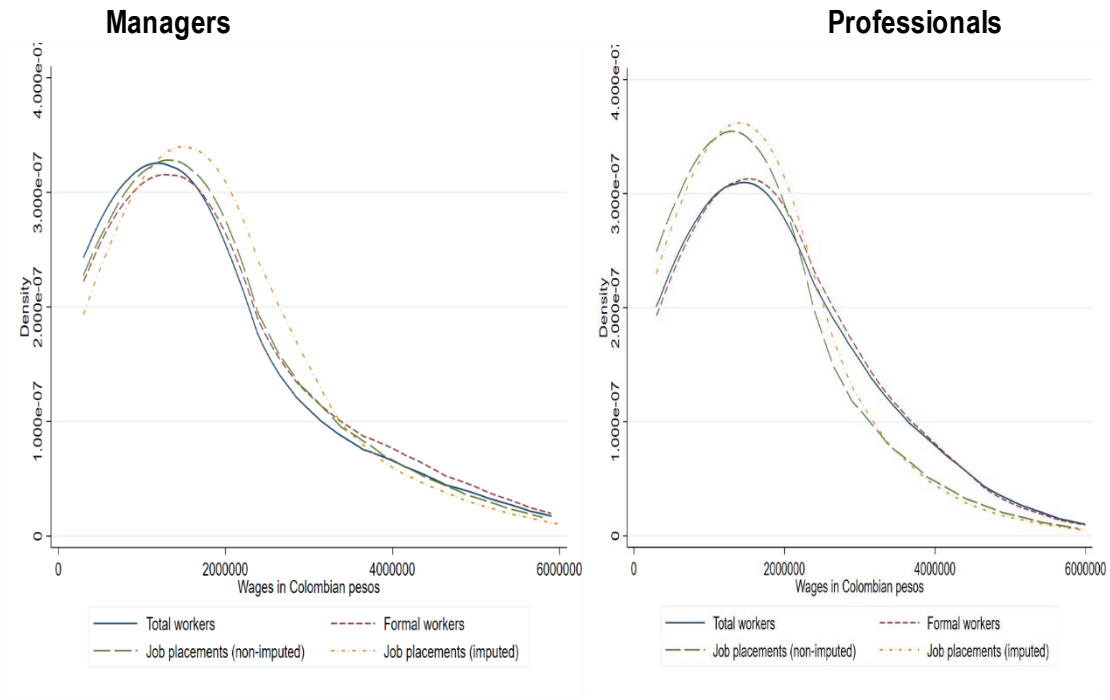
On the one hand, this evidence suggests that the vacancy database does not contain a considerable number of informal jobs; thus, this data might be not representative for the informal sector. On the other hand, formal worker and job portal information (in most cases) have very similar wage distributions. Consequently, the wages in the vacancy database might well represent the "real" salaries that employers are willing to pay for a certain occupation in the formal market, hence job portal information might consistently represent the "real" distribution of vacancies in Colombia.

Third, despite similarities between vacancies and workers' wage distributions, there are some differences. However, it is important to note that more significant differences are found in high-skilled occupations: managers, professionals and technicians, and associate professionals. Banfi and Villena-Roldan (2018) found for the Chilean case that companies tend to post explicit wages when experience or educational requirements are relatively low. Consequently, a company's behaviour might affect the vacancy wage distribution of high-skilled occupations. Indeed, imputed vacancy wages tend to be more on the right tail of the distribution than non-imputed vacancy wages. This result suggests that vacancies with inexplicit wages tend to remunerate their workers more than job advertisements with explicit salaries.

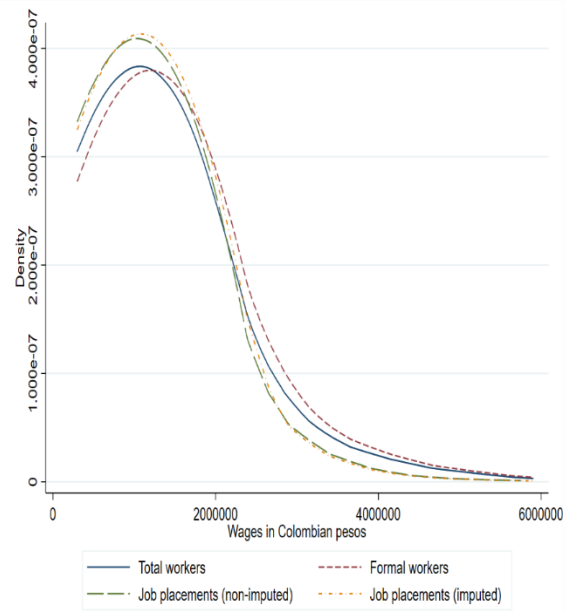
Fourth, the fact that job placements and workers' wages follow similar distributions suggest that "description bias" might not be a predominant issue. These similarities indicate that the workers' and vacancy's salaries are almost the same, hence there are no particular requirements in the job advertisements (e.g. certifications, use of special technologies, etc.) that might increase or decrease wages in the vacancy data and affect their comparison with wages in the labour supply information.

Alternatively, "description bias" might affect the comparison of the vacancy database with informal jobs. As mentioned above, the wage distribution that considers the total number of workers is more to the left than the one that only considers formal workers. These persistent differences might be explained by several reasons. One of them is "description bias"; however, even in this scenario, the differences between informal wage distributions and the vacancy database (for most occupations) are unremarkable, and the shape of the curves are still similar. Thus, at most, the "description bias" affects vacancy data representativeness for the informal sector.

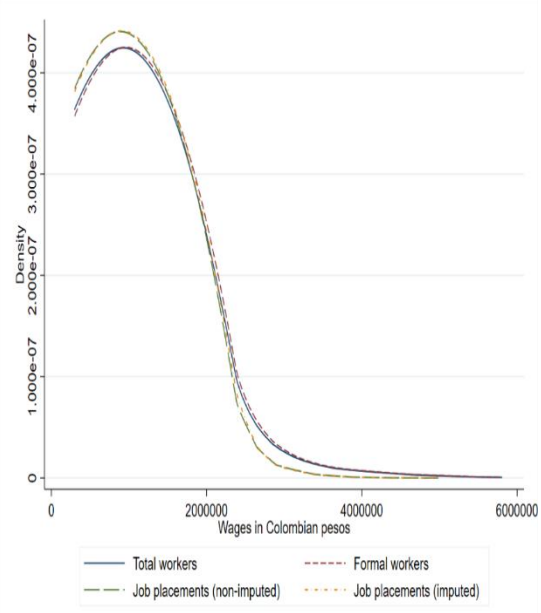
Figure 5. Wage distributions



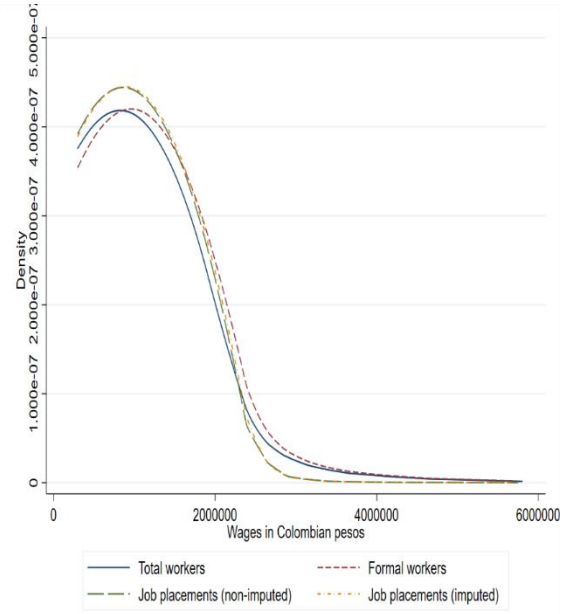
Technicians and associate professionals



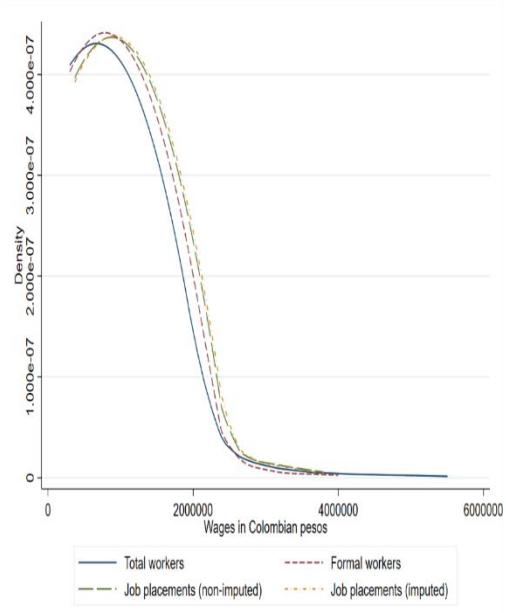
Clerical support workers



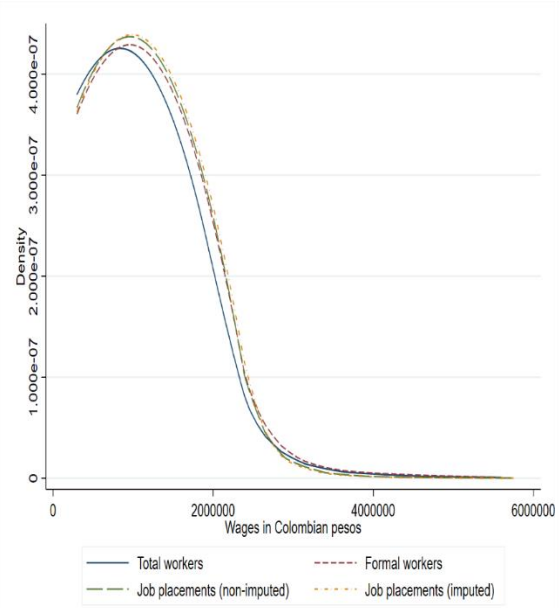
Service and sales workers



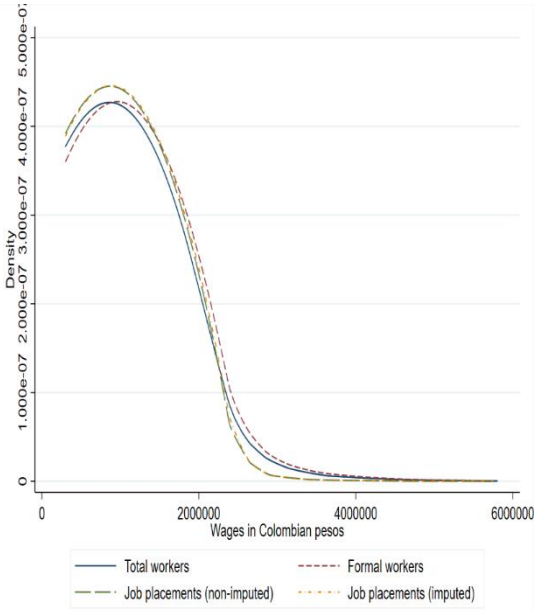
Skilled agricultural, forestry and fishery workers



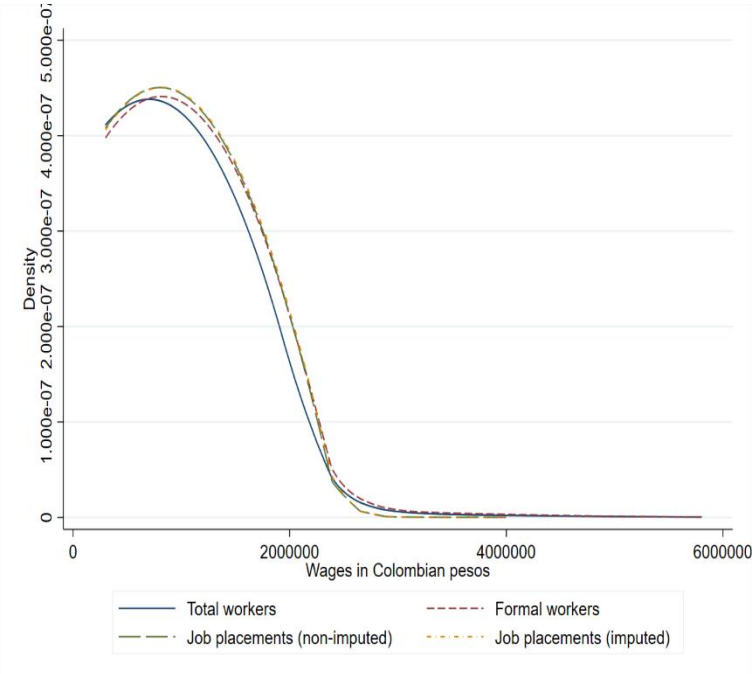
Craft and related trades workers



Plant and machine operators, and assemblers



Elementary occupations



Source: GEIH and vacancy information 2016 - 2018. Own calculations.

As mentioned above, the static-comparative analysis between job placements and workers is limited, although this analysis allows some occupations to be discarded from the vacancy data and provides suggestive evidence regarding the representativeness of the other occupational groups. Nevertheless, given the limitations of this analysis more evidence is needed to validate the data representativeness of the vacancy database.

3.2. Time series comparison

One way to provide further evidence of data representativeness is by comparing labour supply and demand over time ("the labour demand and supply series"). It is not expected that this time series follows exactly the same behaviour because some factors (e.g. skill shortage) might affect the correlation between labour demand and labour series. However, this time series comparison indicates whether economic seasonal and trend effects can be observed in the vacancy database or not. The vacancy database should capture the economic cycles, season and trends to serve as an instrument which informs public policymakers when it is necessary to increase (or decrease) the labour supply of specific skills. However, the period covered by the present study is too short to be certain of anything other than seasonal and (short-term) trend effects.

3.2.1. Stock of people employed

Figure 6 shows the number of vacancies and the number of people employed over time (quarterly from 2016 to 2018) at a one-digit ISCO level (given the large number of occupational groups and the representativeness issues of the GEIH at a four digit-level). The primary axis represents the number of people employed, and the second axis the total number of job placements available in a certain quarter from 2016 to 2018. As can be observed, the series of job placements and people employed follows similar economic seasons for all major occupational groups. Indeed, even the vacancy database follows similar patterns for "Skilled agricultural, forestry and fishery workers". Additionally, the correlation coefficients range from 0.28 for skilled agricultural, forestry and fishery workers to 0.87 for service and sales workers.

This evidence strongly suggests that the vacancy database is a useful instrument to monitor when an occupation is more or less in demand, or whether its demand remains unchanged. Thus, job portal information captures and represents the Colombian economic seasons, and

training provider can potentially use it to estimate when a specific training programme should be increased, decreased, or maintained as it is.

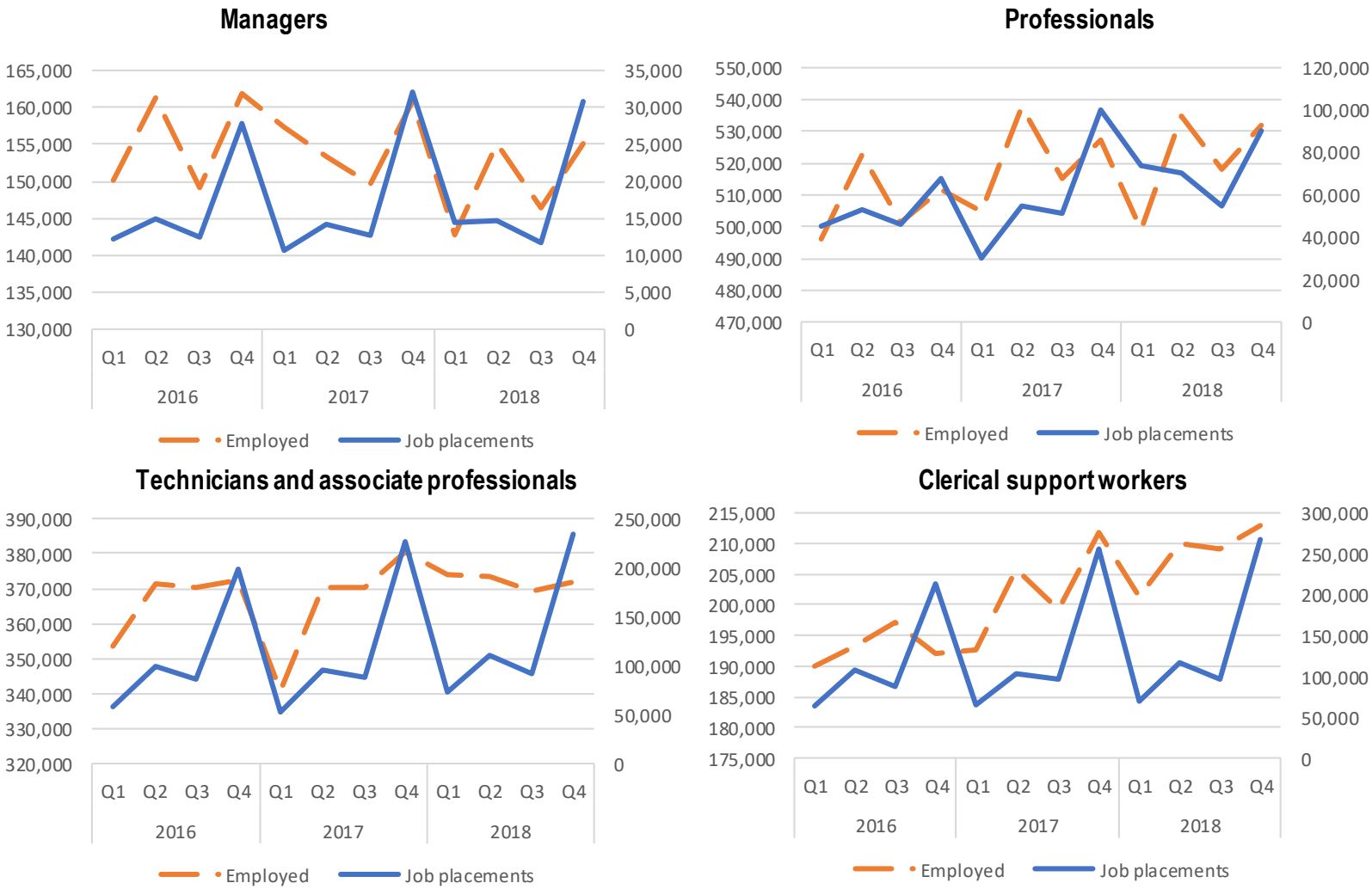
Despite the high correlation between the labour demand and supply series, it is still not possible to determine the exact number of vacancies in the Colombian economy; especially, due to the absence of a vacancy census and the issues mentioned in Cárdenas (2020a). This limitation might affect the labour market and the skill mismatch analysis, specifically, because the employment and the job placement series might increase at the same time. As the exact number of job placements in the market is unknown, a priori it is not possible to know whether the increase in job placements is going to be compensated for by the rise in the number of workers, or not. In this scenario, it would be difficult to determine skill shortages in the labour market.

However, other information available in the vacancy database or the household survey can dispel any doubts regarding whether there are possible skill mismatches. Perhaps, the most useful variable that can confirm the existence of a skill shortage is the wage variable. As noted in Cárdenas (2020a), when a skill mismatch occurs in an occupation or skill, the salaries for that segment of the market start increasing. This and the previous subsections prove the consistency of the wage variable and that economic seasons are reflected in the vacancy database. Consequently, when there is an increase in job placements for specific occupations or skills and, in turn, there is an increase in wages, these circumstances strongly suggest the existence of a skill mismatch. Thus, the vacancy database (at this moment) is not able to provide the exact or approximate number of job placements, yet the information can be used to identify possible skill shortages (See Cárdenas, 2020b).

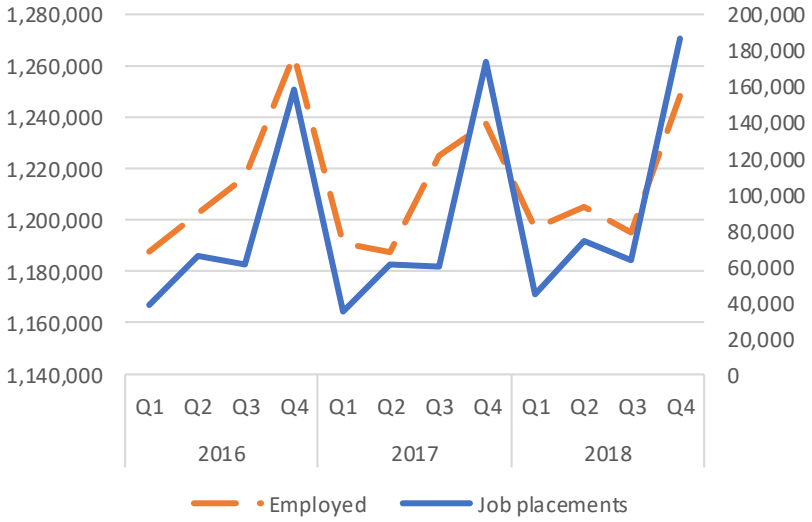
Moreover, the total number of vacancies in the economy can be, potentially, estimated. As mentioned in Cárdenas (2020a), labour demand is comprised of both the level of employment (satisfied labour demand) and the number of available job vacancies which denote the labour not filled by an employee over a certain period (unsatisfied labour demand or unmet demand). In turn, the unmet demand is calculated from the separation rate (the total number of employees who left their jobs) and the total of new jobs created. By estimating the separation rate, the job destruction rate, and sectoral and occupational employment growth rates, similar to Flórez et al. (2017), it might be possible to estimate the level of unmet labour demand and contrast it with the

vacancy database. However, the calculation of these parameters will be part of future work, given the complexity of this task.

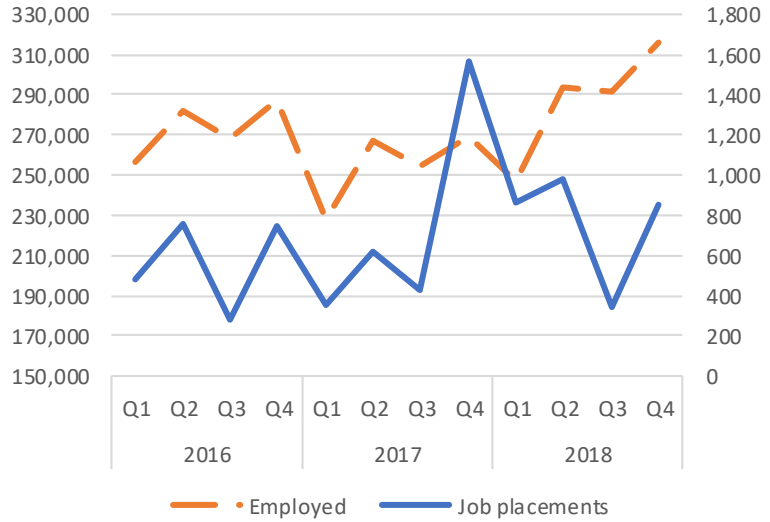
Figure 6. Time series: total employment and job placements 2016–2018



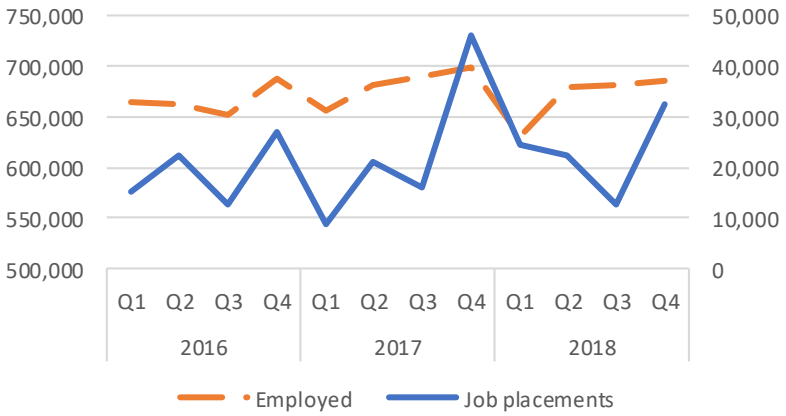
Service and sales workers



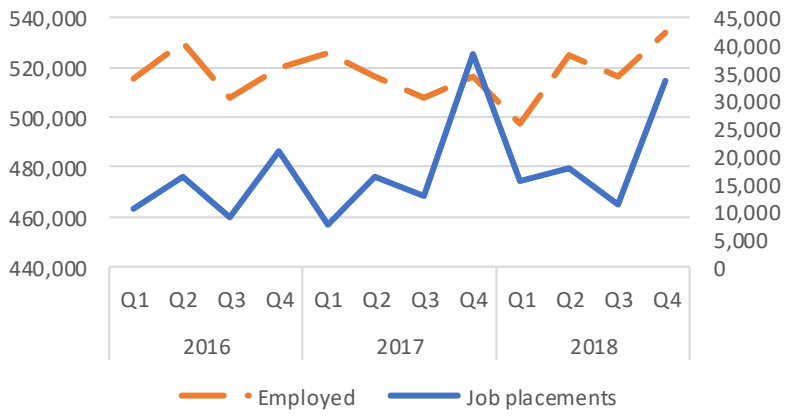
Skilled agricultural, forestry and fishery workers

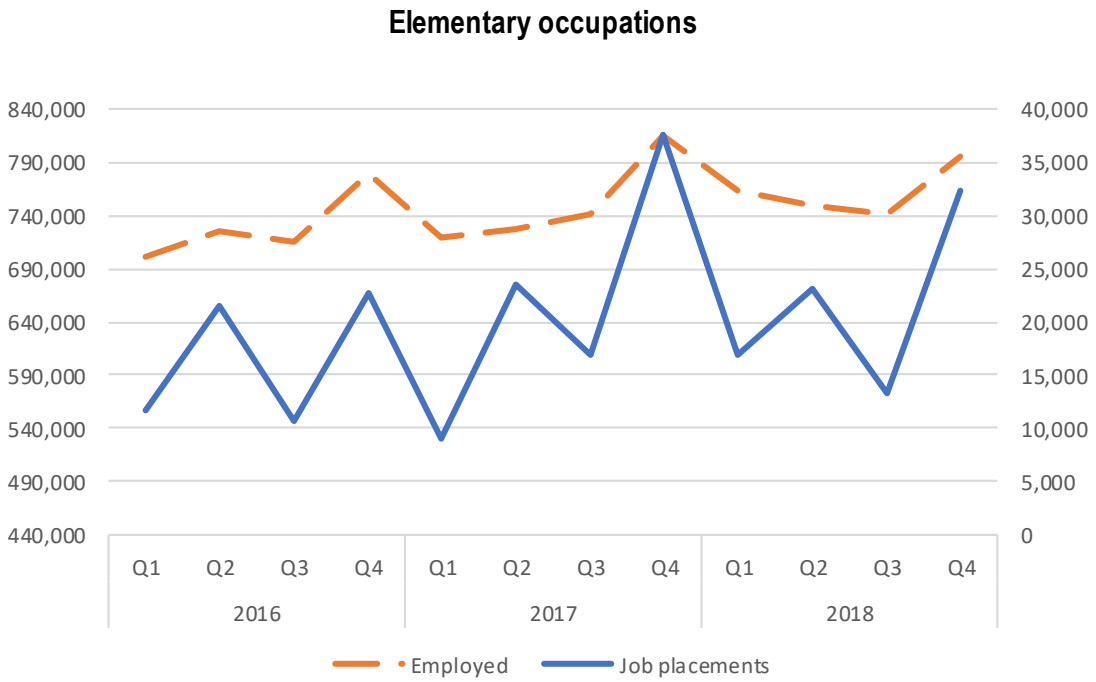


Craft and related trades workers



Plant and machine operators, and assemblers





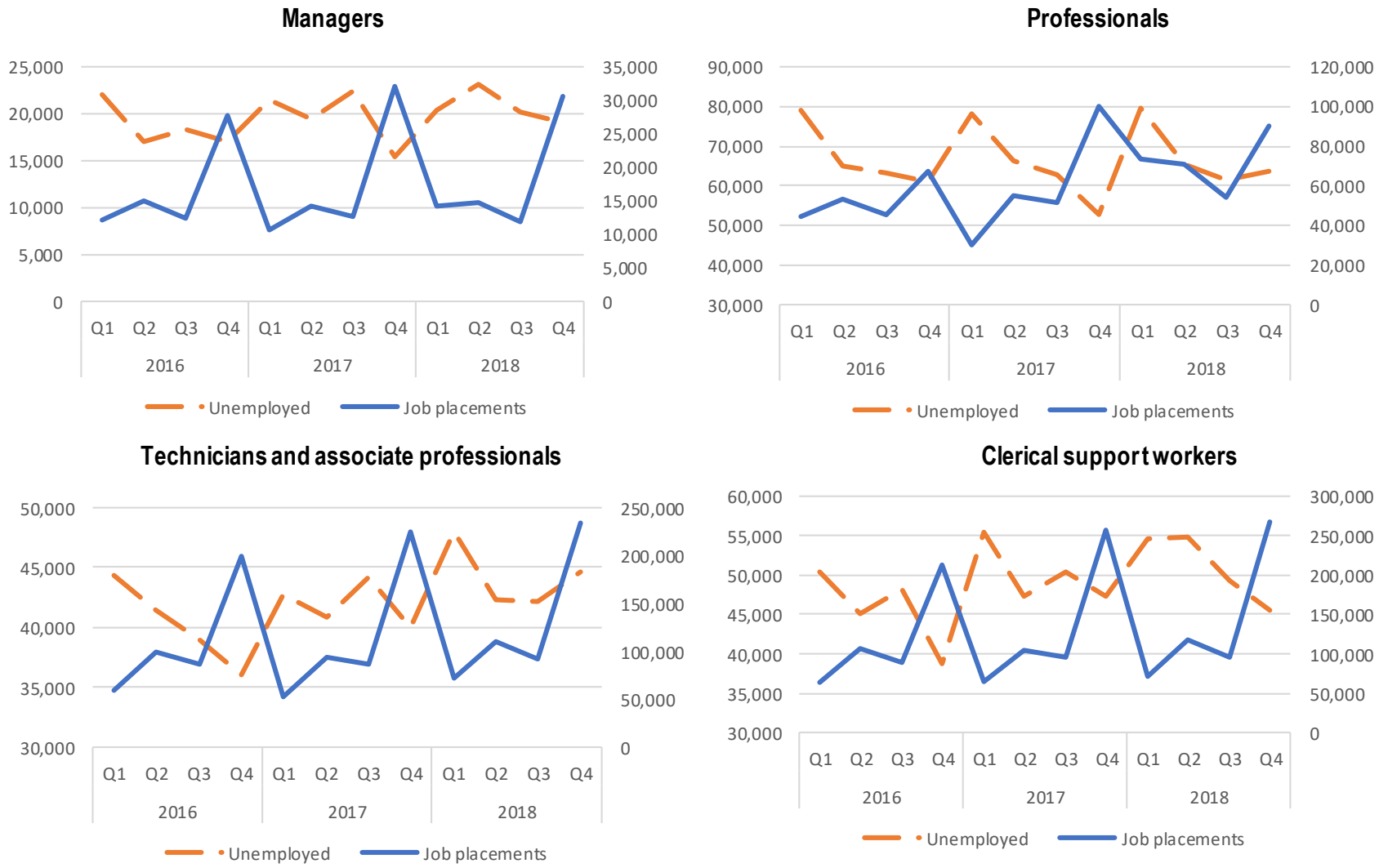
Source: GEIH and vacancy information. Own calculations.

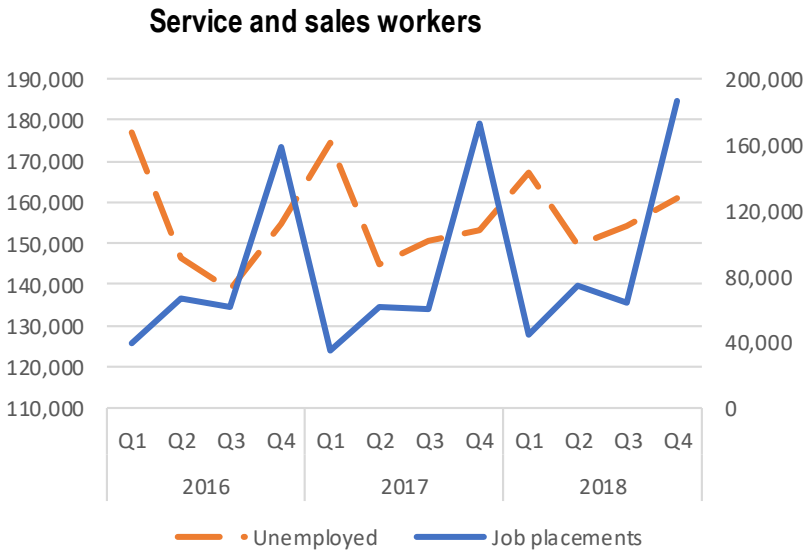
3.2.2. Stock of people unemployed

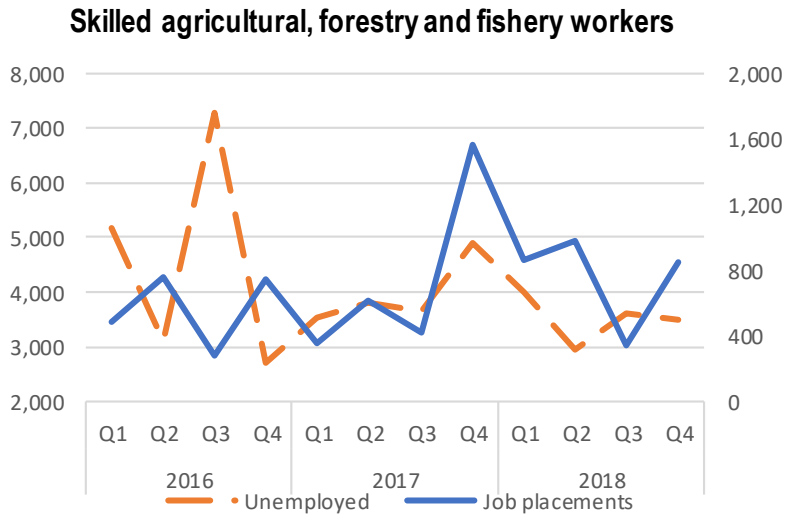
The above comparison showed that the vacancy database has a strong correlation with employment rates in Colombia. To provide more evidence regarding the external consistency of the information gathered from job portals, and to demonstrate that vacancy data can be used to build different labour market indicators, this subsection compares the vacancy series with the level of unemployment. Usually, periods of high unemployment are associated with low levels of vacancies and vice versa (e.g. the Beveridge curve, see Cárdenas, 2020b).

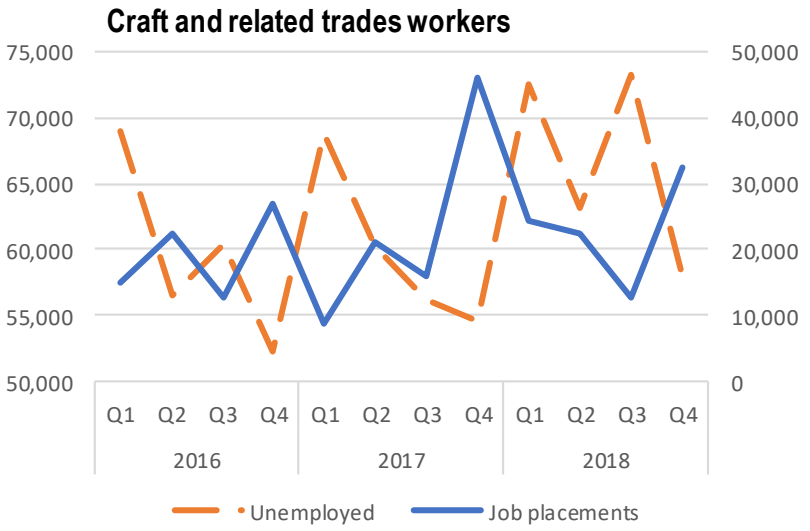
Figure 7 shows a time series to compare unemployment figures against the number of job placements. As expected, in general, these series are negatively correlated for all occupational groups; demonstrating that when there is an increase in the number of job placements the level of unemployment decreases. The correlation coefficients range from -0.15 for "Service and sales workers" to -0.65 for "Managers". Thus, the results from the vacancy database are consistent with the unemployment series from the official survey. Moreover, these results suggest that it is possible to combine vacancy information with the unemployment level to build indicators to monitor the labour market, such as the Beveridge curve, by occupational groups (see Cárdenas, 2020b).

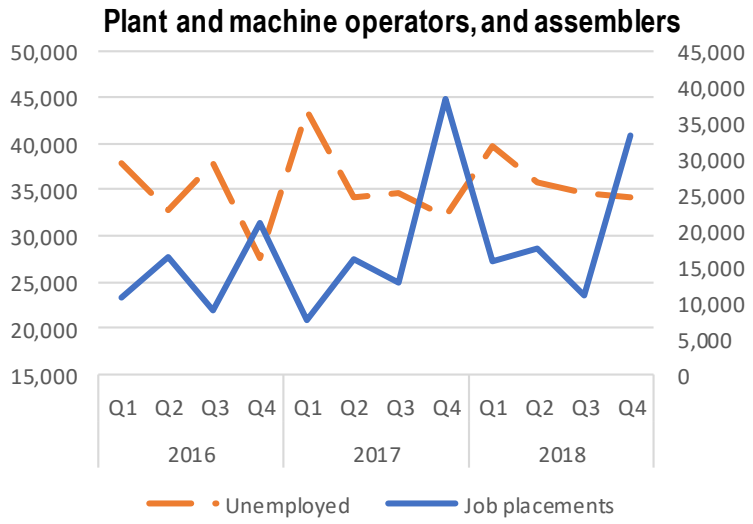
Figure 7. Time series: total unemployment and job placements 2016–2018

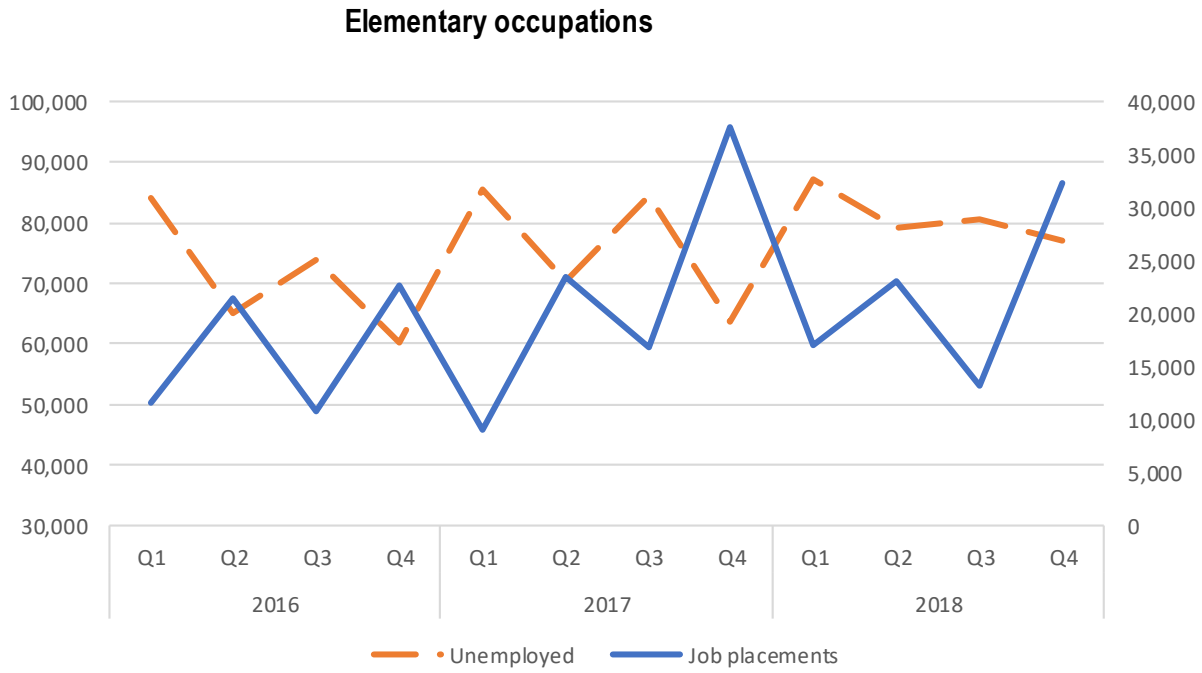



Service and sales workers
190,000
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
200,000
160,000
120,000
80,000
40,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016
2017
2018
Unemployed
Job placements


Skilled agricultural, forestry and fishery workers
8,000
7,000
6,000
5,000
4,000
3,000
2,000
2,000
1,600
1,200
800
400
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016
2017
2018
Unemployed
Job placements


Craft and related trades workers
75,000
70,000
65,000
60,000
55,000
50,000
50,000
40,000
30,000
20,000
10,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016
2017
2018
Unemployed
Job placements


Plant and machine operators, and assemblers
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016
2017
2018
Unemployed
Job placements



Source: GEIH and vacancy information. Own calculations.

3.2.3. New hires (replacement demand and employment growth)

As mentioned above, the comparison between the total workforce and job placements is the most common way to test the data representativeness of the vacancy database. However, this exercise might be limited. The total workforce is composed of the total number of employed and unemployed people, while job portals contain information regarding the net and replacement labour demand (see Cárdenas, 2020a) (Lmiforall, 2018). The total workforce is a measure of the labour market "stock", while the number of job vacancies is a measure of the labour market "flow". Consequently, the similarities (or dissimilarities) between the workforce and the job placements time series might be due to other labour dynamics such as participation or dismissal rates, rather than a causal effect between the number of vacancies and the number of employed or unemployed people.

For instance, the last subsections showed that positive correlation occurs between the number of job placements and the number of employed people; especially, in the last quarter of each year the number of employed people and the number of job placements are relatively higher. However, this correlation might be due to a lower dismissal rate. Assuming, at the very least, that real opening jobs rates are consistent in each quarter of the year, it might happen that in the last quarter of the year dismissal rates are relatively lower than the other quarters because employers need to keep more workers for the Christmas season, and thus the number of employed people is higher. Consequently, the vacancy data collected from job portals might not correctly represent the dynamics of real job openings, even when there seems to be a high correlation with the employment and unemployment series.

To test this argument, it is necessary to compare the vacancy series with the net growth[13] plus the replacement demand[14]. It is not possible (so far) in Colombia to identify the total number of vacancies, and much less to distinguish the net growth and replacement demand separately. However, with the Colombian household survey information, it is possible to know when people started working. Specifically, the GEIH asks the following question: "How long has [interviewed

[13] Net growth refers to the number of job openings as a consequence of economic growth or decline,

[14] Replacement demand refers to the number of job openings created because of people changing employers, occupations, sector, etc., as well as people temporally leaving their jobs (e.g. sickness), retirement or death.

name] been working in this company, business, industry, office, firm or farm continuously?". With this question, it is possible to estimate the number of people who start working in the previous months (new hires). In other words, the number of new hires (which fills vacancies) created by economic growth (net growth), and the number of vacancies created because people left their jobs (replacement demand). Consequently, new hires have a strong correlation with the number of job openings and, thus, if the vacancy database properly represents the dynamics of job openings, the vacancy data should be correlated with the new hires time series.

It is important to note that new hires do not entirely represent labour demand. As mentioned above, the household survey provides information regarding the number of job matches. Consequently, new hires are signified by the net growth plus the replacement demand matched in the previous months. Nevertheless, there is no strong reason to think that the new hires (matched) time series are not correlated with the number of vacancies available. One argument might be that vacancies occur for certain occupations, but there are no people with the skills and (other) characteristics required. Therefore, vacancies can be created but not (necessarily) new hires. This argument might be valid for a detailed labour market analysis (e.g. at a four-digit ISCO level). However, general trends and seasonal information for the new hires at an aggregated level (e.g. at a one or two-digit ISCO level) should be reflected in the household survey.

Otherwise, in the Colombian labour market there are huge barriers such as skill mismatches that prevent people being hired even when there is an increase of vacancies at the occupational group level (at a one or two-digit level). Nevertheless, and as mentioned above, this argument does not seem plausible because if there is such an evident barrier to match jobs, the economy and the government would react to correct the issue without the need of a detailed labour market analysis.

Figure 8 depicts the number of new hires and job placements in a quarterly time series[15]. These time series comparisons show an important fact: the new hires and the job placements have a strong lagged correlation. Indeed, when time series are compared within the same period the Pearson correlation coefficient is between -0.68 and 0.04, and when the new hires are lagged by one period (one quarter) the Pearson correlation coefficients sit between 0.17 and 0.70

[15] Given GEIH representativeness issues the data is quarterly aggregated.

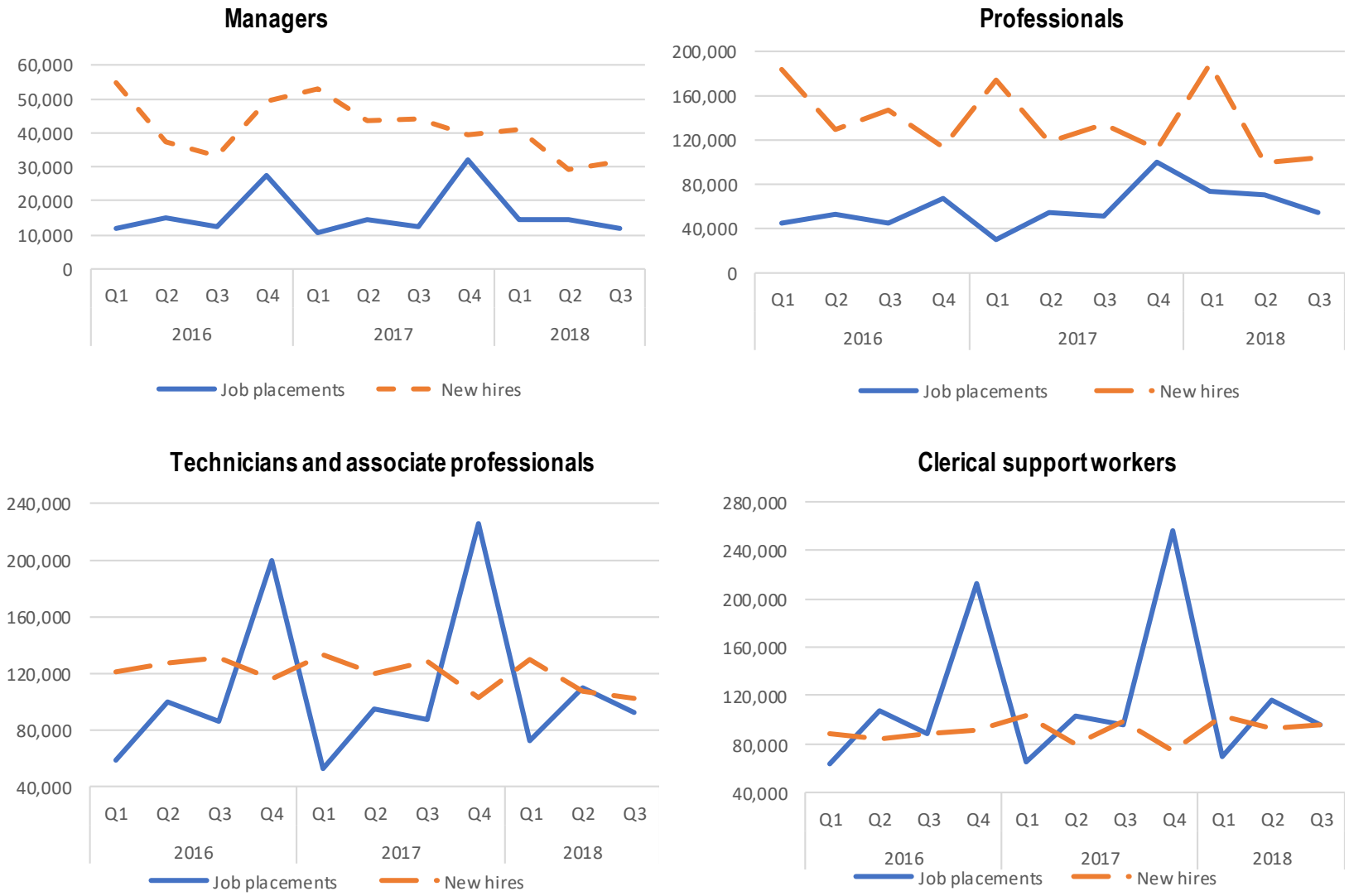
(except for “Skilled agricultural, forestry and fishery workers” whose correlation coefficient is -0.01).

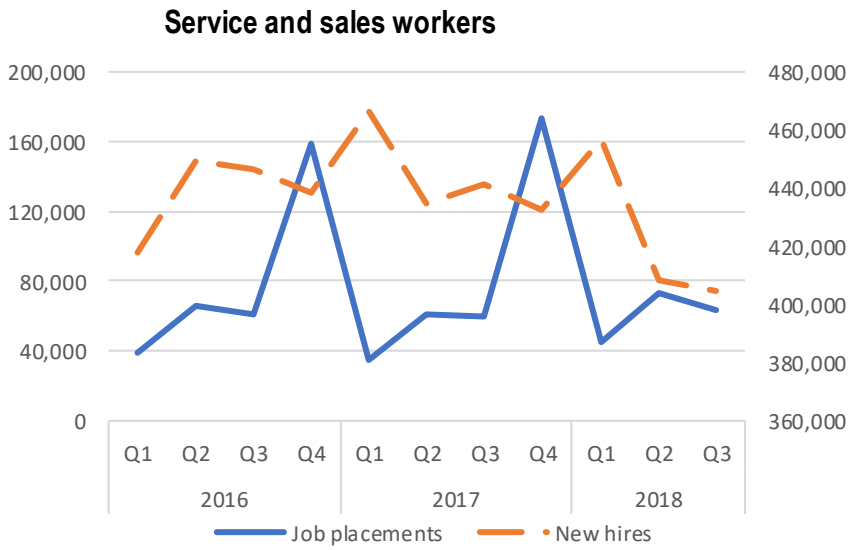
These results suggest that there is a lagged effect between the increases and decreases of job placement advertisements and the number of people who occupy these job positions. As mentioned in Cárdenas (2020a), posting vacancies is part of the search process, and one of the first steps taken to hire workers. Between posting the vacancy and hiring the most appropriate worker requires time and effort for both the employee and the employers (indeed the median duration of advertising is 1.2 months—see Cárdenas, 2020d). Companies need to attract a certain number of workers, after which companies carry out screening, selecting, and training, among other processes, while workers need to surmount all those processes and, in some cases, work a period of notice with their existing employer.

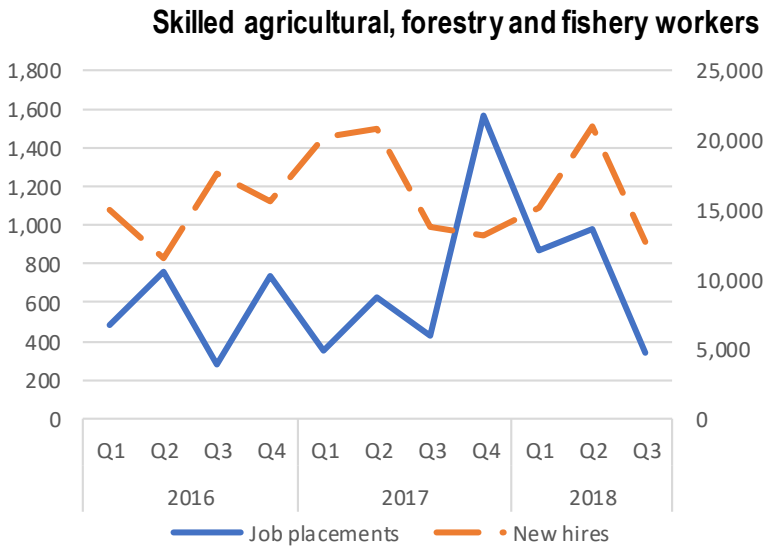
Thus, a lagged correlation is expected between increases/decreases of job advertisements and the moment when people occupy these jobs. Moreover, this lagged correlation shows the dynamics and timing of the hiring process in Colombia. For instance, Cárdenas (2020d) showed that for all occupational groups the number of job advertisements sharply increases between October and November, which makes sense given that, as Table 4 describes below, November is the third month when there are additional hires (8.5% of new hires occur in this month during 2016 and 2018).

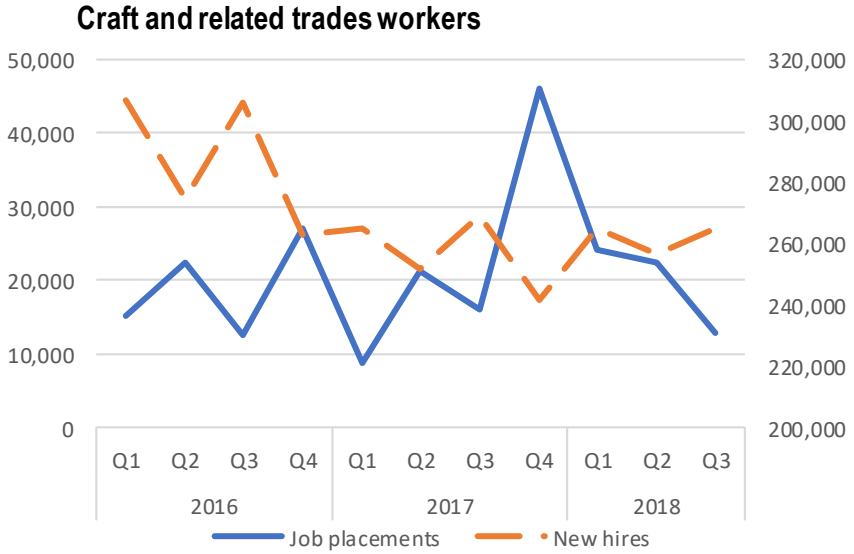
However, overall January is the month in which there relatively more new hires. This behaviour is because in November companies start hiring people for the Christmas season (see Cárdenas, 2020d). Nevertheless, in December new hires usually decrease because in this month a considerable portion of people are on vacation (in Colombia, December is well-known as a period where students and most workers take relatively long vacations). Consequently, hiring processes are usually slow in December. On the contrary, January is the start of the new fiscal year when companies become more active again and hire a portion of those people who were contacted and selected in the previous months. This evidence suggests that trends and economic seasons for new hires are strongly correlated with the number of job advertisements, hence the vacancy database adequately represents these trends and the economic season of the total number of job placements.

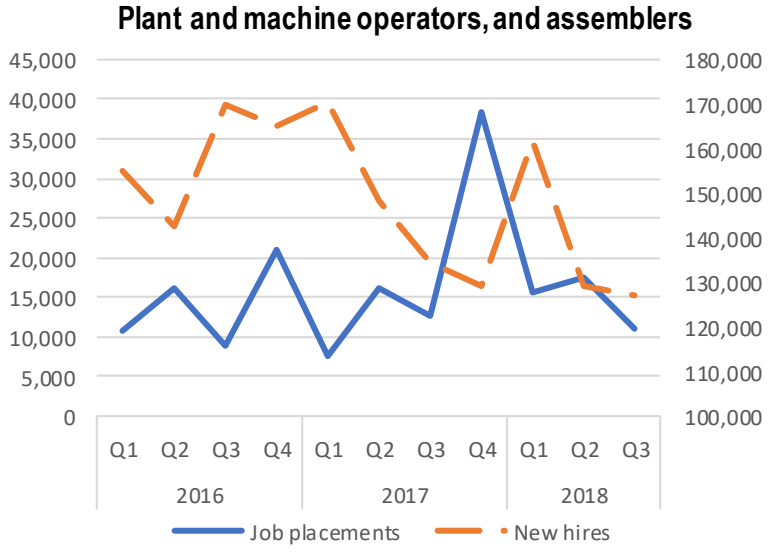
Figure 8. Time series: new hires and job placements 2016–2018

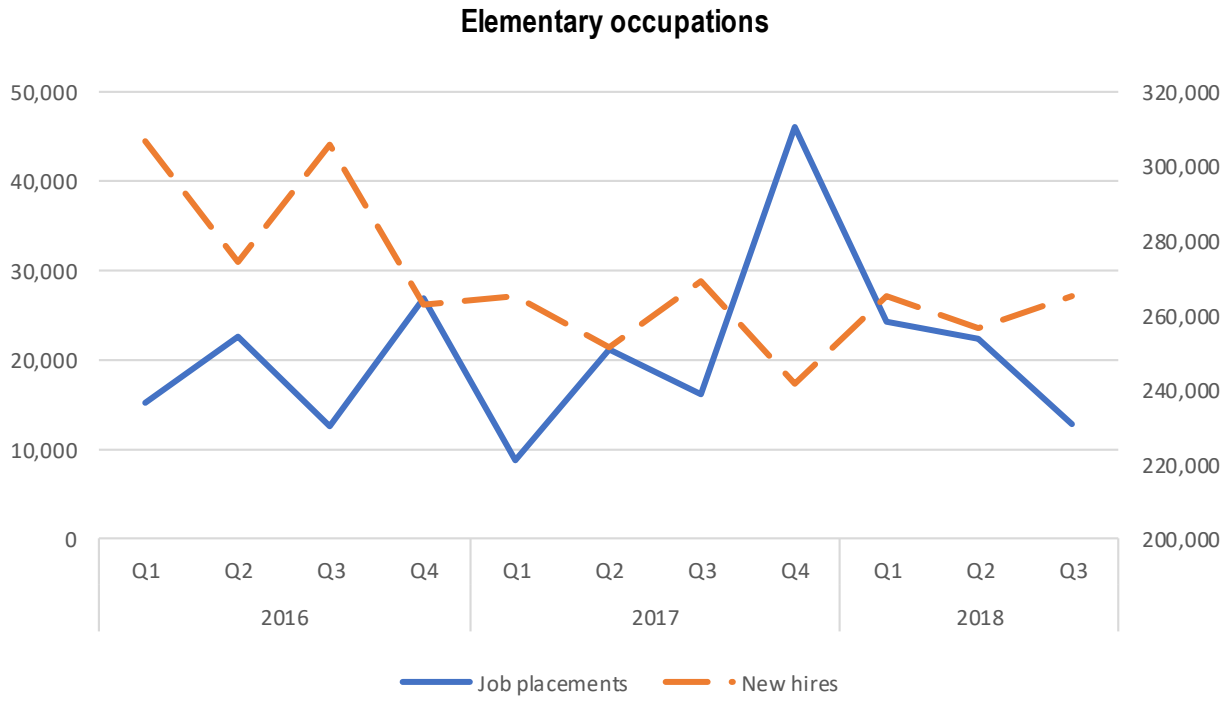



Service and sales workers
200,000
160,000
120,000
80,000
40,000
0
480,000
460,000
440,000
420,000
400,000
380,000
360,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Skilled agricultural, forestry and fishery workers
1,800
1,600
1,400
1,200
1,000
800
600
400
200
0
25,000
20,000
15,000
10,000
5,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Craft and related trades workers
50,000
40,000
30,000
20,000
10,000
0
320,000
300,000
280,000
260,000
240,000
220,000
200,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Plant and machine operators, and assemblers
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
100,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires



Source: GEIH and vacancy information. Own calculations.

Table 4. Monthly distribution of new hires 2016–2018

Month	Percent
January	10.4%
August	8.7%
November	8.6%
July	8.5%
September	8.5%
February	8.5%
October	8.4%
June	8.4%
March	8.2%
May	8.1%
April	8.1%
December	5.8%

Source: GEIH information 2016 - 2018. Own calculations.

Consequently, the evidence suggests that for the Colombian case, the vacancy database provides (per se) meaningful information about skills and employers' requirements. In general, the occupational structure of the vacancy information at a four-digit ISCO level is coherent with the information from official surveys, especially, for the urban economy, formal and non-agricultural occupations. The seasonal and economic trends for a considerable share of the labour market are captured at least at a one-digit ISCO level. Moreover, this data combined with wage, employment and unemployment information can potentially warn policymakers, educators and workers about potential skill shortages.

4. Conclusion

Any database has limitations. To test the validity of the database's information is a paramount process to avoid misinterpretation and biases in the analysis. In the case of the vacancy database, which is composed of online job advertisements, different concerns arise (see Cárdenas, 2020a). For instance, information from the Internet might not correlate with general characteristics of the labour market, or the algorithms that collect and organise job advertisements might fail. Consequently, this paper provided an evaluation of the internal and external consistency of the vacancy results.

On the one hand, internal validity refers to the consistency of the variables within the vacancy database (Henson, 2001; Streiner, 2003); in that, the results from a variable in the vacancy database should not contradict the findings from other variables in the same data. The findings of this test show that the contradictory or inconsistent results occurring in the Colombian vacancy database were minor, and the magnitude of these measurement errors are insufficient to bias the educational, occupational, sectorial, skills and wage analyses.

On the other hand, external validity refers to the consistency of the results from the vacancy database when compared with information from other sources (in other words data representativeness) (Rasmussen, 2008; Stopher, 2012). The vacancy data per se can provide valuable answers about what people should be trained in at a low cost (time and money). Nevertheless, testing the data "selection bias" of the vacancy database is challenging because of the absence of a vacancy census, or any official data that supplies the total number of vacancies in Colombia (the statistical universe).

Despite the different difficulties, this paper provided an external evaluation utilising sources of information available in the country. Thus, a static comparison was made between labour supply and vacancy information. First, the occupational structure of the vacancy database (labour demand) and the GEIH (labour supply) was compared. This comparison provided three conclusions: 1) the vacancy database is not representative for a significant part of agricultural, government and armed force occupations; 2) particular caution should be taken when analysing occupations with high turnover rates as this issue might cause an overrepresentation of specific occupational groups; and, 3) self-employed individuals ("business owners") and informal occupations are not represented in the vacancy database. This evidence suggests that the vacancy database better represents the formal and urban Colombian labour market.

Second, a comparison between the distribution of wages in the vacancy database and the GEIH was carried out. This exercise suggests that wages in the vacancy database well-represent the "real" salaries that employers are willing to pay for a particular occupation, and the comparison also shows that the vacancy database might consistently represent the distribution of vacancies in Colombia.

Moreover, the vacancy database should capture economic seasons, cycles and trends to serve as an instrument which can inform public policymakers when it is necessary to increase (or decrease) the labour supply of specific skills. Consequently, a number of time series comparisons between the number of vacancies and people employed, unemployed and new hires were made to establish whether economic seasons could be observed in the vacancy database or not. This comparison showed that job portal information captures and represents the Colombian economic seasons. In general, when the level of job placements increases, so does the level of employment; conversely, when there is an increase in the number of job placements, the level of unemployment decreases. Importantly, the comparison between new hires and the job placements revealed that the trends and economic seasons for new hires are strongly (lagged) correlated with the number of job advertainments, hence the vacancy database adequately represents the "real" trends and economic seasons of the total number of job placements. Thus, training providers could potentially use the vacancy database information to estimate when training provision should be increased, decreased or maintained. However, so far, economic cycles could not be analysed because of the relatively short period of information available from the database (three years).

It is not possible (at this moment) to determine the exact number of vacancies in the Colombian economy, mainly, because of the absence of a vacancy census. However, it is not necessary to comprehend a precise amount of vacancies in the economy to identify possible skill shortages, among other essential characteristics of the labour market. A rigorous analysis using information from the online job portal vacancies and GEIH data (such as wages, trends, occupational structure, etc.) provide sufficient information to design indicators (such the Beveridge curve, or wage and employment trends) and determine possible skill shortages for a significant segment of the Colombian labour market.

Thus, the vacancy database, in general, is representative of a considerable set of formal, non-agricultural, non-governmental, non-military and non-self-employed ("business owners") occupations over 2016 to 2018. Despite the fact that the vacancy information does not capture

a considerable share of agricultural jobs, the relatively few observations in the vacancy database for those occupations might provide insights to policymakers, educators and workers about new skill requirements and general trends for some agricultural occupations.

5. References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings.Handbook of labor economics, 4, 1043-1171.

Backhaus, K. B. (2004). An exploration of corporate recruitment descriptions on Monster. com. Journal of Business Communication, 41(2), 115-136.

Banfi, S., & Villena-Roldan, B. (2018). *Do high-wage jobs attract more applicants? Directed search evidence from the online labor market*. Journal of Labor Economics, Forthcoming.

Cárdenas R., Jeisson. (2020a). Information Problem in Labour Market and Big Data: Colombian Case. Universidad del Rosario. Working, paper No. WP2-2020-001.

Cárdenas R., Jeisson. (2020b). Possible uses of labour demand and supply information to reduce skill mismatches. Universidad del Rosario. Working, paper No. WP2-2020-002.

Cárdenas R., Jeisson. (2020c). Extracting value from job vacancy information. Universidad del Rosario. Working, paper No. WP2-2020-003.

Cárdenas R., Jeisson. (2020d) Descriptive analysis of the vacancy database. Universidad del Rosario. Working Paper No. WP2-2020-004.

Carnevale, A. P., Jayasundera, T., & Repnikov, D. (2014). *Understanding online job ads data: a technical report.* Georgetown University, McCourt School on Public Policy, Center on Education and the Workforce, April.

Flórez, L. A., Morales, L. F., Medina, D., & Lobo, J. (2017). *Labour flows across firms size, economic sectors and wages: evidence from employer-employee linked panel.* Borradores de Economía, (1013).

Henson, Robin K. (2001). *Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. (Methods, plainly speaking)*. Measurement and Evaluation in Counseling and Development, vol. 34, no. 3, 2001, p. 177+. Academic OneFile, Accessed 29 Feb. 2019.

Kennan, M. A., Willard, P., Cecez-Kecmanovic, D., & Wilson, C. S. (2008). *IS knowledge and skills sought by employers: Analysis of Australian IS early career online job advertisements.* Australasian Journal of Information Systems, 15(2).

Kureková, L. M., Beblavy, M., & Thum, A. E. (2014). *Using internet data to analyse the labour market: a methodological enquiry* (No. 8555). IZA Discussion Papers.

Kureková, L. M., Beblavy, M., Haita, C., & Thum, A. E. (2016). *Employers' skill preferences across Europe: between cognitive and non-cognitive skills.* Journal of Education and Work, 29(6), 662-687.

LinkedIn (2019) *These Are the 5 Types of Jobs with the Most Turnover*. [online] Available at: https://business.linkedin.com/talent-solutions/blog/talent-analytics/2018/these-are-the-5-types-of-jobs-with-the-most-turnover [Accessed 13 Mar. 2019].

Lmiforall (2018). *What is replacement demand?* [online] Available at: http://www.lmiforall.org.uk/2017/05/what-is-replacement-demand/ [Accessed 23 Apr. 2018].

OECD (2017c). *OECD Employment Outlook 2017.* OECD Publishing, Paris. http://dx.doi.org/10.1787/empl_outlook-2017-en

Rasmussen, K. B. (2008). *General approaches to data quality and Internet-generated data.* The Sage handbook of online research methods, 79-97.

Štefánik, M. (2012). *Internet Job Search Data as a Possible Source of Information on Skills Demand (with Results for Slovak University Graduates).* In Building on Skills Forecasts — Comparing Methods and Applications, edited by CEDEFOP. Luxembourg: Publications Office of the European Union. http://www.cedefop.europa.eu/EN/Files/5518_en.pdf

Stopher, P. (2012). *Collecting, managing, and assessing data using sample surveys.* Cambridge University Press.

Streiner, D. L. (2003). *Starting at the beginning: an introduction to coefficient alpha and internal consistency*. Journal of personality assessment, 80(1), 99-103.

Agradecimientos

Esta serie de documentos de trabajo es financiada por el programa "Inclusión productiva y social: programas y políticas para la promoción de una economía formal", código 60185, que conforma Colombia Científica-Alianza EFI, bajo el Contrato de Recuperación Contingente No.FP44842-220-2018.

Acknowledgments

This working paper series is funded by the Colombia Científica-Alianza EFI Research Program, with code 60185 and contract number FP44842-220-2018, funded by The World Bank through the call Scientific Ecosystems, managed by the Colombian Ministry of Science, Technology and Innovation.