Extracting value from job vacancy information

Jeisson Cárdenas

ALIANZAEFI
economía formal e inclusiva

Documento de Trabajo
Alianza EFI - Colombia Científica
Marzo 2020

Número de serie: WP2-2020-003

Extracting value from job vacancy information[1] [2]

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

Abstract

This paper presents a comprehensive methodology to collect and standardise vacancy information systematically from job portals. Describes available information in Colombian job portals. Describes the methodology (web scraping) and challenges to automatically and rapidly collect a massive number of online job vacancies. Also explains the methods that can be used to homogenise variables, and explains challenges involved in standardising two of the most relevant variables for the economic analysis of the labour market: skills and occupations. This paper develops a method to automatically identify skills patterns in job vacancy descriptions based on international skill descriptors and text mining. In addition, it conducts a novel mixed-method approach (software classifiers and machine learning algorithms) to properly classify job titles into occupations. Furthermore, it deals with duplication and missing value issues, by using predictors such as occupation, city, and experience requirements.

Key words: Big Data, web scraping, text mining, machine learning, skills, occupations

JEL classification: C88, J23

[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. Methodology

1.1 Introduction

The analysis of labour demand information is a relevant factor to improve people's skills according to employers' requirements. As mentioned by OECD (2017) the capacity of countries to improve and adjust their labour supply according to labour demand for skills determines different labour outcomes such as productivity and economic growth, among others, and in the context of this thesis, unemployment, informality, etc. However, as discussed in Cárdenas (2020a), this capacity to analyse the labour demand, in most countries, has been hampered by a lack of information about employers' requirements.

Recently, online job portals have caught the attention of researchers and policymakers insofar as they might fill the labour demand information gap (Kureková et al. 2014; Reimsbach-Kounatze, 2015). These job portals contain a large number of job adverts which are accessible to anyone interested in vacancies and employers' requirements. Despite this information being publicly available, the analysis of labour demand using job portals is challenging. First, there are large numbers of job advertisements available, online dispersed over different websites; consequently, there is not a consolidated database to use to analyse labour demand information. Second, each job portal manages information according to their own criteria. For instance, some websites might use the term "wage" while others use "salary", or some websites might show remuneration information with numbers while others display them with words or ranges (e.g. monthly £2,000, or two thousand pounds per month, or between £1,750–£2,250 monthly). Moreover, relevant information such as job titles or demanded skills are not categorised to facilitate labour demand analysis.

For the reasons mentioned above, the vacancy information is not organised, categorised and consolidated in a database for statistical purposes. Thus, it is necessary to develop a robust methodology which collects, organises, categorises and analyses labour demand using job portals. This section proposes and explains each of these methodological steps. The second subsection of this section describes what information is available from Colombian job portals. The third subsection analyses the most important and reliable job portals to investigate how to conduct a proper labour demand analysis. Given that there are a large number of vacancies

available online and, consequently, the manual collection of labour demand information is virtually impossible, the fourth subsection describes web scraping techniques that can be used to automatically collect online job advertisements. The fifth subsection explains the organisation (homogenisation) of different job portal information into a single database the once the information is collected. Specifically, it explains how programmed algorithms search the information of each vacancy for patterns to build education, experience, localisation and wage variables from the text. However, not all the variables in the vacancy database can be built using the same method (looking for textual patterns in job advertisements). For instance, to build a variable such as "company sector" is necessary to implement other and more complex text mining techniques; thus, the last subsection of this section shows how it is possible to identify the sector where the employer belongs.

1.2 Measurement of the labour demand: job vacancies

As mentioned in more detail in Cárdenas (2020a), a job vacancy can be understood as a vacant position within a company that the company is trying to fill. Companies recruit potential workers in diverse ways to fill their vacancies. Likewise, as discussed in Cárdenas (2020a), job portals provide companies with an informatics platform to make public the number and characteristics of available job positions over a certain period. Even though job portals are not the only channel where companies advertise their vacancies (for instance, occupations related to IT tend to be over-represented, see Cárdenas 2020a), they might capture a large share of the net and replacement labour demand behaviour.

Table 1.1 shows the most important job portals in terms of data traffic (the number of visitors) available in Colombia (Alexa, 2017)[3]. For instance, "https://www.jobportal_a.com.co/"[4] is the 37th most visited web page in Colombia, while "https://www.jobportal_b.com/" is 89th. Additionally, Column 3 in Table 1.1 shows the number of job advertisements available for each job portal in October 2017.

[3] Alexa Internet, Inc., is a wholly owned subsidiary of Amazon.com which calculates and ranks the data traffic of a website based on the browsing behaviour of the Internet users of each country.

[4] This Working Paper anonymised the name and web page address of job portals to protect the identity, and not promote a particular job portal.

A job advertisement is understood as text on a job portal which shows relevant information about a job vacancy (Nigel, 2016), and a single job advertisement can contain one or more job vacancies (i.e. Mass recruitment). Consequently, the first thing to note from Table 1.1 is that there are a large number of job advertisements and job vacancies on each website[5]. This amount of data makes the manual collection of information a task that would require many working hours and/or a large number of people employed in a monotonous task; that is, to copy the information and paste it in a database thousands of times.

Table 1.1: Average number of job advertisements and traffic ranking for selective Colombian job portals

Colombia	Alexa Rank	Number of Job Adverts
https://www.jobportal_a.com.co/	37	115,723
https://www.jobportal_b.com/	89	62,732
https://www.jobportal_c.gov.co/	199	263,621
https://www.jobportal_d.com.co/	1,015	172,440
https://www.jobportal_e.com.co/	2,280	20,143
https://www.jobportal_f.com.co/	3,683	46,853

Source: https://www.alexa.com and the job portals

Each job portal shows a list of available vacancies. Nevertheless, each of the websites organises and shows its data according to their own criteria (see Appendix A:). Table 1.2 (below) summarises the difference between two job advertisements within the same job portal. Even though this website presents almost the same information between the two vacancies, the localisation and categorisation of these variables (such as experience and wages, among others) might vary according to website design and the information provided by the employer or recruitment agency. Moreover, some job advertisements on the same website might contain more or less information than the example listed in Table 1.2. Consequently, a job portal is a semi-structured source of vacancy information. This feature makes it difficult to collect data from these website sources automatically. Thus, an algorithm that collects this information needs to recognise differences between advertisements and organise the information to properly

[5] In Colombia, companies advertise their vacancies on different websites and, depending on the job portal, the cost of promoting a vacancy varies between £24 to £26.

construct/calculate totals for the net and replacement labour demand database (hereinafter labour demand database—see Cárdenas 2020c).

Table 1.2: Job advertisement structure comparison within the same job portal

Variables	Panel A: First job advertisement				Panel B: Second job advertisement			
	Box A	Box B	Box C	Box D	Box A	Box B	Box C	Box D
Job title	X			X	X			X
Experience	X					X	X	
Wage	X			X		X		X
Location	X			X	X	X		X
Publication date	X				X			
Company name								X
Description		X				X		
Number of jobs		X				X		
Education requirement			X				X	
Type of contract				X				
Workday				X				
Age required							X	

Source: https://www.jobportal_a.com.co

Differences between job announcements also arise when comparing two different websites. For instance, Figure 1.1 compares two job advertisements: one posted on Jobportal_a (Panel A) and the other posted on Jobportal_c (Panel B). Both adverts required an "accountant" (see Box A, Panel A and Panel B). Note that these are not the same vacancy posted on different websites. However, the information is displayed in a different way. For Jobportal_a, information about job requirements (such as education, experience, etc.) and job characteristics (such as wage, type of contract, etc.) are shown in the C Box (at the bottom of the Panel A) and the D Box (on the right of Panel A). In contrast, Jobportal_c displays information about job requirements and job characteristics together in Box B (on the left of Panel B).

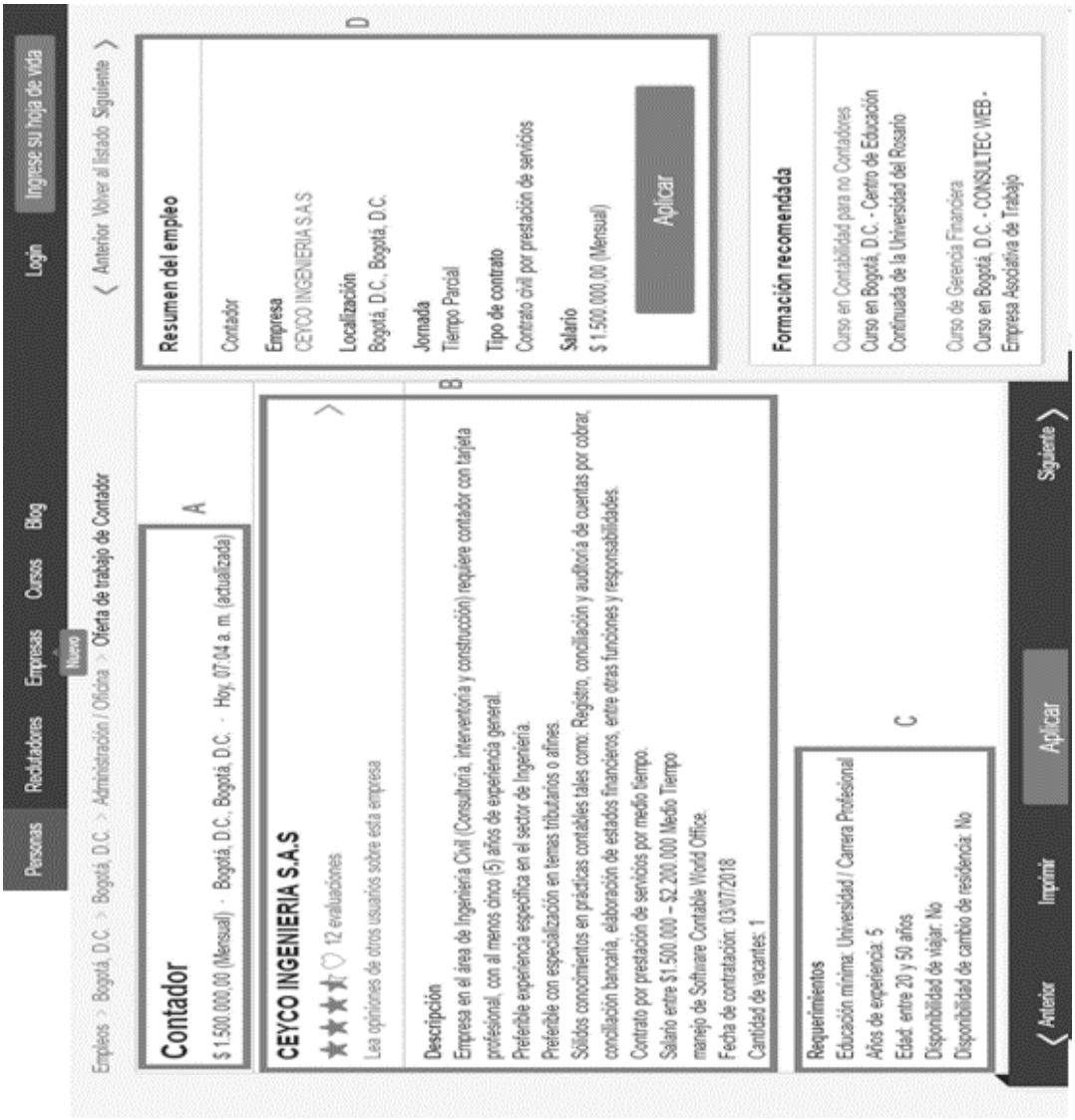
Additionally, variables such as wages or experience might be categorised in different ways. On Jobportal_a wages are expressed in numbers (in this case 1,500,000 Colombian pesos monthly), and the experience requirement is expressed in terms of years. In contrast, for

Jobportal_c the wage variable is expressed in ranges based on the official minimum wage[6]. This minimum wage, fixed by the government each year, acts as a reference point for employers. For instance, a vacancy might offer a salary between "1 a 2 SMMLV" (which means between one and two times the minimum wage), while another vacancy might offer a wage between "two and four times SMMLV" (which means 2 to 4 the minimum wage); and the experience variable is shown in terms of months for Jobportal_c.

[6] In Colombia, every year, the national government decree the minimum remuneration for a full-time job. For the year 2018, the minimum wage was 781,242 Colombian pesos (around £196) per month.

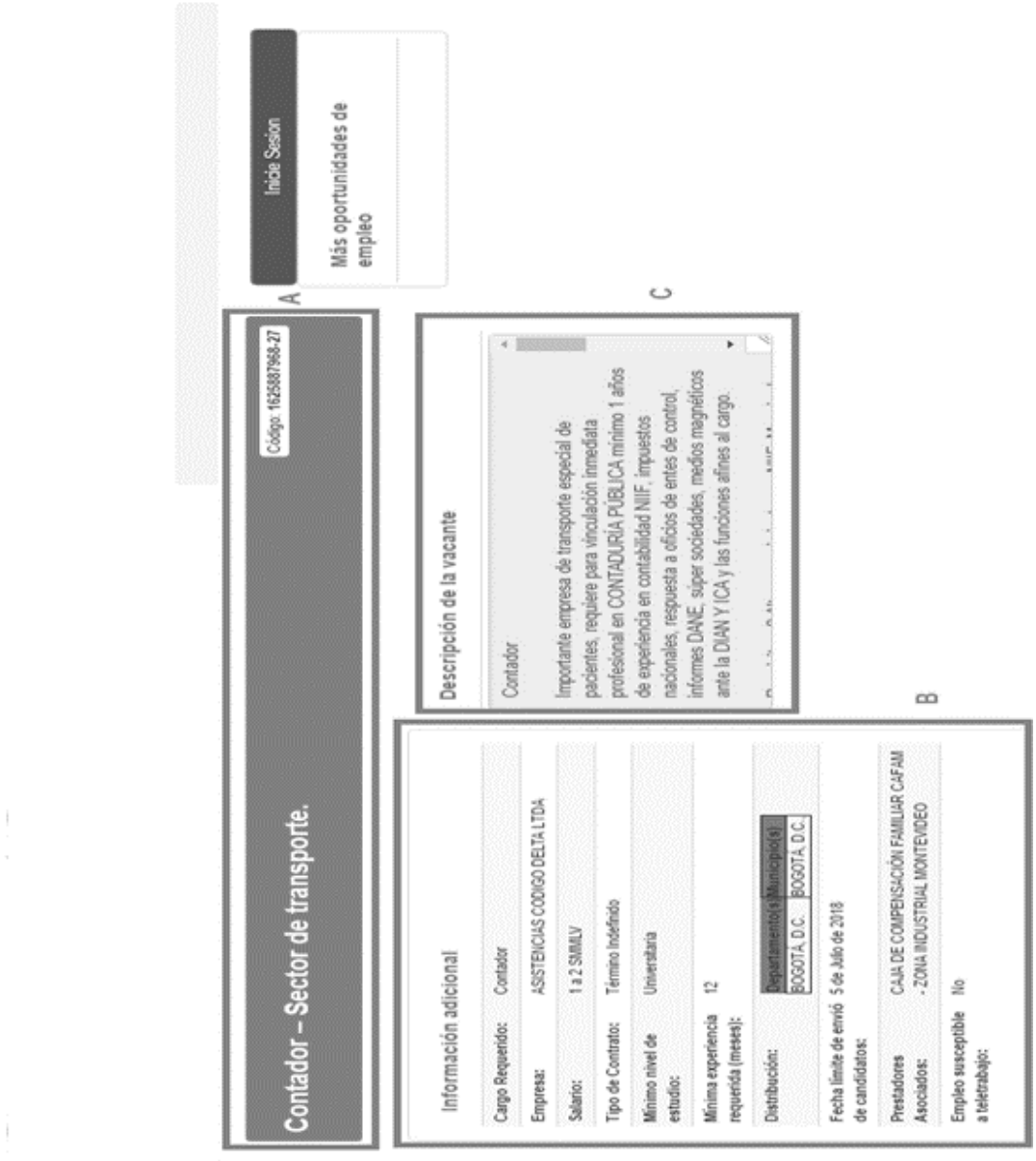
Figure 1.1: Job advertisement comparison between job portals

Panel A: Jobportal_a[7]



[7] Box A stands for: Accountant. Wage 1,500,000 pesos (monthly). City: Bogotá D.C. Department: Bogotá D.C. Posted: Today at 07:04 am; Box B: Company's name: CEYCO Ingenieria S.A.S. Description: Accountant is required, with at least five years of general experience. Contract of service: Part-time. Accounting software: Word office. Date of hire: 03/07/2018. The number of jobs: 1. Box C: Requirements. Minimum Undergraduate certificate. No travel is required. Five years of work experience. Age: 20 to 50 years old. Box D: Job summary: Accountant. Company's name: CEYCO Ingenieria S.A.S. Localisation: Bogotá D.C.. Working day: Part-time. Type of contract: Contract of service. Wage: 1'500,000 pesos (monthly).

Panel B: Buscadordeempleo[8]



Source: https://www.jobportal_a.com.co and buscadordeempleo.gov.co/

[8] Box A stands for: Accountant, Transport sector. Box B: Accountant. Company's name: Asistencias codigo detal LTDA. Wage: from 1 to 2 SMMLV. Indefinite term contract. Minimum undergraduate certificate. Twelve months of work experience. City: Bogotá D.C. Expiry date: 5th July 2018. Box C: Accountant. Minimum 1 year of experience in accounting NIF, national taxes, among others.

Even though these format and structural differences might be regarded as superficial to the human eye, they represent a challenge for the automatic collection of labour demand information. First, the structural differences between job portals correspond to differences in how each website was programmed. Specifically, websites can be programmed in different programming languages—such as HTML (HyperText Markup Language), Javascript, PHP (Hypertext Preprocessor), ASP (Active Server Pages), and so forth—and these languages can be integrated (e.g. an HTML code might contain a JavaScript code). Each of these programming languages possesses its own structure and functions (see Appendix A:, Figure A.3)[9].

This heterogeneity between and within websites makes it difficult to collect information automatically. For each job portal, it is necessary to develop an algorithm that recognises the programming language, the structure, and can extract the relevant information from each website and each job announcement. Thus, in order to collect as much information on labour demand as possible, the first part of my methodology involves the following steps:

- Select the most important vacancy websites in the country.
- Scrape the vacancy websites selected.
- Apply text and data mining techniques to organise the information.

1.3 Selecting the most important vacancy websites in the country

As shown above, different websites exist with relatively high data traffic (a high number of visitors) and with a significant volume of job advertisements. However, there are a variety of issues to consider before extracting information from job portals. Firstly, there is a trade-off between the number of job portals and the time/effort required to build a vacancy database: as more portals are considered an increase in effort (human and computational capabilities) and time investment is needed to program each algorithm for each job portal. Additionally, the structure of websites might change over time and, consequently, algorithms need to be adjusted accordingly to those changes, and the effort and time to collect information from websites increases significantly as a result.

[9] For instance, in HTML information is delimited by tags, such as “<img/>”, “<a>”, etc., while information in JavaScript language uses syntax such as “<script type="text/javascript">” “</script>”.

Secondly, when considering a larger number of portals duplication problems arise (as discussed in section 2 in more detail). Companies or recruitment agencies might post the same vacancy on different job portals. As a consequence, the use of many websites increases the probability of duplication. Even though this problem can be diminished by different techniques (see section 2), the probability of duplication persists and increases by adding more websites. Yet, if a single job portal is used to build a vacancy database other issues arise[10]. Results derived from that website might be biased or limited in their representativeness of the overall job market. Therefore, in terms of obtaining a certain level of quantity (representativeness) and quality the selection of job portals is a critical stage in the building of a vacancy database.

Provided that relevant sources of job vacancy information and computational capabilities exist, to decrease the possible bias of utilising one source it is necessary to consider the job adverts from different websites to build a vacancy database. In order to select the job portals that best capture the dynamic of the Colombian labour market, the following criteria were applied to select job portals: 1) volume (the number of advertisements available), 2) website quality (structure and number of variables or granularity of information), and, 3) traffic ranking (total number of users). Regarding the former, as shown in Table 1.1, job portals that seemed to have more vacancy information were Jobportal_c (263,621 job vacancies), Jobportal_d (172,440 job vacancies) and Jobportal_a (115,723 job vacancies).

However, the volume of posted information should not be the only element to select the most relevant job portals. First, some job portals might post a job advertisement that was originally posted on other job portals. Such is the case for Jobportal_c and Jobportal_d [11]. Consequently,

[10] For instance, a job portal might be focused only on a specific segment of the market (e.g. graduate or IT jobs).

[11] Jobportal_d announced that the website had a total of 172,440 job vacancies available on 30th October 2017. However, when clicking on some vacancy announcements the new window displayed, gave a brief and short description of the vacancy and provided the link where that vacancy was originally posted and where an interested person might find more information regarding the job. Similarly, Jobportal_c announced that the website had a total of 263,621 job vacancies available on 30th October 2017. However, when clicking on some vacancy announcements, the new window displayed redirected the search and opened another website where the vacancy was posted (e.g. Jobportal_a).

websites such as Jobportal_c and Jobportal_d do not necessarily contain a major number of job advertisements[12].

Moreover, the amount of information (the number of advertisements) is not the only factor that matters to select the best job portals and to build a vacancy database. The degree of detailed information provided by each website is another element to be considered in the selection process. The more detailed the information, the better the inputs are to build variables such as skills, wages, education, etc. Thus, the second criterion to select a job portal is the granularity of information provided by the job portals. In this sense, except for Jobportal_d, the job portals listed in Table 1.1 show similar variables on their websites. Indeed, to post a vacancy on these websites, the employer needs to supply a minimum of information (required fields). This guarantees, with some minor variations, that these job portals, usually, have information regarding the job title, city, wages offered, education requirements and the company name, among others. In contrast, the Jobportal_d does not have a pre-defined format where employers need to fill the corresponding information. To post a vacancy on this website it is only necessary to complete the job title, and employers might or might not provide more detailed information in the vacancy description. Therefore, to consider a job portal such as Jobportal_d might increase the number of cases with missing values in the vacancy database.

The third criterion to select job portals is the number of users measured by the website's traffic ranking. The number of users might indicate individuals' (companies and job seekers) "trust" regarding the information provided on a particular website. Additionally, to take in to account the traffic ranking of websites, to some extent, guarantees that the selected sites do not specialise in a specific category of vacancies, such as graduate or IT jobs (see Cárdenas 2020b for more evidence regarding this point). As Table 1.1 shows, Jobportal_a, Jobportal_b and Jobportal_c. are the websites that have a higher number of visitors.

Consequently, after an exploration of job portals based on the three elements mentioned above, I have selected the following web pages to be scraped and analysed because they have a relatively high number of job announcements (volume), users (traffic) and are well-defined websites (quality):

[12] The magnitude of this redirect issue was unknown at this stage of the methodology.

Table 1.3: Job portals and main characteristics

Job portal	Main characteristics
Jobportal_a	It is a widespread private platform in Latin America[13]. In Colombia, this source is third in terms of the number of observations (vacancies) posted, it has a minimum number of requirements fields (semi-organised), and it is the most used job portal in Colombia.
Jobportal_b	It is a private platform that operates in Colombia, Costa Rica, Peru, Guatemala and Salvador. In Colombia, this source is fourth in terms of the number of observations (Colombian vacancies), it has a minimum number of requirements fields (semi-organised), and it is the second most used job portal in Colombia.
Jobportal_c	It is a platform administrated by the Colombian Government (more specifically by the Unidad del Servicio Público de Empleo: UAESPE). This source is first in terms of the number of observations (vacancies), it has a minimum number of requirements fields (semi-organised), and it is the third most used job portal in Colombia.

Moreover, as the Alexa Rank shows in Table 1.1, there is a significant difference between the rank of the first three job portals (Jobportal_a, Jobportal_b and Jobportal_c) and the remainder (Jobportal_e and Jobportal_f). Additionally, a manual check showed that Jobportal_e and Jobportal_f are not specialist websites that cover job types not found on the three selected job portals. This evidence suggests that reliable information on the total number of vacancies in Colombia might be concentrated in the first three job portals, those visited the most according to their Alexa Rank (Cárdenas 2020b demonstrates that the job portals selected offer a variety of jobs from low-skilled to high-skilled positions).

Finally, it is important to note that the quality and quantity of information provided by the sources mentioned above might change over time. Moreover, platforms that were not taken into account or new ones might start providing valuable information (increasing the number of advertisements, increasing the number of users, etc.). This dynamic might change which job portals should be considered for the construction of a future vacancy database(s). Thus, the

[13] Indeed, there is a version of this platform for: Colombia, Peru, Argentina, Uruguay, Guatemala, Ecuador y El Salvador, Honduras, Venezuela, Nicaragua, Cuba y Costa Rica, Mexico, Chile, Panamá, Dominican Republic, Bolivia, Paraguay and Puerto Rico.

evaluation of job portals should be a constant process to guarantee that the best sources of information are selected to provide the best possible labour demand information.

1.4 Web scraping

As was previously seen in subsection 1.2, the differences between and within job portals require differences in programming language and codification structure. Hence, to obtain and analyse labour demand information in Colombia I implemented a technique called "web scraping" which consists of a computerised method to automatically collect information from across the Internet (in this case from vacancy portals) (oxforddictionaries, 2017). Broadly speaking, this is attained through different software that simulates human web surfing to collect specified parts of public information (job advertisements) from various websites, and store them in a database to be further organised and analysed.

Although the information is not adequately organised to identify each variable, websites have labels, headers, nodes, tags, among other markers, within their HTML code, that allow the extraction of the most relevant information within the data. Codes in R software were built to make this automatic collection of the data possible. With the codes that this thesis develops, the computer is programmed to visit each job advertisement announcement, to copy all relevant information related to the description of the vacancies, and to paste it in a unique database to be organised and analysed. The codes should be built in such way that the computer recognises each of the job portal's structures, auto-adjusts the number of vacancies to be scrapped, and automatically subtracts and saves the relevant information, among other processes and rules. Thus, to program the codes knowledge is required in HTML, CSS and programming language such as R (see Appendix A:, Figure A.3).

Since each web portal displays vacancies in a semi-structured way, they do not follow a well-defined standard to display the data: the Xpaths[14] change between one website and another. Moreover, so far, there is not an automatic way to determine which Xpaths contain the relevant vacancy information. As a consequence, the selection of Xpaths needs to be done manually for each website. This selection process requires a certain knowledge of HTML programming language to select the information correctly. Given the difference between the HTML structure

[14] An is an expression used to identify nodes in websites.

from one website compared with another, it is necessary to create a different code for each web portal in order to download the relevant vacancy information. In consequence, for this thesis this method required the construction of three different codes: one for Jobportal_a, one for Jobportal_b and the other one for Jobportal_c [15].

Once the codes are created, the next step is running the programs to download the corresponding information[16]. Each time the codes are run, the (uncleaned) data is saved in a (local) personal server. Importantly, information downloads should be checked periodically. Job portals might inadvertently change their HTML structure. As a consequence, codes might become outdated and fail to extract vacancy information. In this case, the corresponding codes should be updated according to changes in the website structure. However, if there is a long gap between a significant change in the HTML structure of a website and the update of the corresponding code, this might represent an unrecoverable loss of information over a certain time period[17].

[15] The scraping of each website requires different packages and software. While scraping websites such as Jobportal_a and Jobportal_c does not require sending security credentials (e.g. a login via a user account) to have access to the information, other websites such as Jobportal_b request a login and the sending of other user's credentials. This login issue (among other issues) makes it necessary to connect R with a software-testing framework such as "selenium" for the scraping of websites such as Elmpleo. Thus, the codes and computing tools (packages and software) to scrape information from job portals might differ significantly between job portals.

[16] The process of downloading data using web scraping for a website such as Jobportal_a can last one day, meaning that the computer visits around 80 announcements per minute to obtain the information required. While extracting information from a website such as Jobportal_b takes around three days. These time differences depend on factors such as the web page response time of each job portal, the maximum number of connections allowed, internet speed, sending user credentials, among other factors.

[17] For instance, consider a job portal which has 50 vacancies in October 2017, and the corresponding code failed to obtain that information due to changes in the website. In November 2017 the same job portal has 100 vacancies available, 80 of which are new vacancies while 20 correspond to vacancies published in the previous month (October). Thus, in November 30 vacancies that were published in October 2017 are not available any more on the website (the jobs were filled and/or the employer paid to post the vacancy for a short period). Consequently, if the code is updated in November 2017, 30 observations from October and their information would have been lost (if the vacancy links are dropped or unavailable on the website).

Therefore, first, it is critical to periodically review (via a visual inspection) that each of the codes is extracting the corresponding information, and, second, to run the codes frequently to avoid significant information loss between one download and the next. For this thesis, each code was run three times per month to avoid information loss.

1.5 The organisation and homogenisation of information

Once the data is obtained, the next step is to provide a well-defined structure to the semi-structured data collected from the vacancy portal. As seen in Appendix A:, the localisation (XPath) of a variable might change between job adverts. XPath changes might cause some columns in the database to be out of line. For instance, a column that should correspond to education might contain information about job experience, and vice versa.

Since the information on online jobs boards is semi-structured, it is necessary to use natural language processing techniques to organise vacancy information. Specifically, it is required to use methods to analyse unstructured data such as word analysis (text mining) in order to obtain unified variables, such as wages, work experience, education level, geographic area, and the skills required by employers.

1.5.1 Education, experience, localisation, among other job characteristics

First, it is necessary to carry out a reading of a set of job advertisements to identify the keywords that employers use to describe the characteristics of job positions (such as experience, type of contract, localisation and education). Once keywords are identified, an algorithm is written in order to "read" the job vacancies which generates a dummy variable that takes the value of 1 if a particular pattern is in a job advertisement (see Appendix B:).

Not all variables can be classified into dummies variables, however, given the multiple values that some variables can take, which is the case for localisation, wage, company name and occupational variables that can accommodate many values, such as the names of different cities, towns, a salary in numbers or words, etc. For this reason, the implementation of another text mining process is required in order to organise and homogenise this vacancy information.

1.5.2 Wages

Employers might or might not provide wage information in job advertisements. When they provide this information it can take different forms, e.g., wages might be expressed in numbers or words. Moreover, job portals such as Jobportal_b display wage information according to a minimum and maximum range. For instance, a vacancy might contain the following information regarding the wage offered: "$1.5 a $2 millones mensuales"[18]. Given the diverse forms that wage information might take, I followed a number of steps. First, I programmed an algorithm that searches and extracts wage information (in whatever form it takes) from job advertisements[19].

Second, once the information was extracted and placed in a single column it was necessary to apply a homogenisation process. As mentioned above, wage information might be displayed in diverse forms. For those cases where the wage revealed the exact number of pesos that a worker would receive once hired, I did not apply any depuration, but where wages were described in words I transformed the words into their equivalent in numbers. Additionally, when wages were shown in ranges, I selected the average value between the maximum and the minimum range. It is important to note that in the above steps, I looked for explicit information about wages and imputation procedures were not yet implemented (section 2 will discuss the issues regarding missing values and possible ways to handle them).

1.5.3 The classification of companies

The labour demand for skills is produced by a group of private and public companies that perform different activities and provide goods and services. Depending on those activities and goods and services, companies are classified into sectors. Evidently, the skills required from one industry to another sector might differ. Sectors such as mining tend to ask for people with knowledge in controlling heavy machinery for the exploitation of underground mines, while the Information and communication sector tends to require people with knowledge in programming. Moreover, there are some generic skills such as communicating and problem-solving that might be used in different sectors. Thus, the analysis of vacancies by sector might identify which skills or

[18] Around £375–£400 monthly.

[19] The usage of the peso ($) symbol or the word "pesos" (which is the Colombian currency) aided the identification of information regarding wages.

occupations are sector-specific or generic. Addressing labour supply according to the needs of each sector, might reduce skill mismatches in the labour market by providing skill information to educators, training institutions and policymakers (see Cárdenas 2020a).

Frequently, job portals provide information about the company that is advertising a job position. Part of this information might be useful when identifying the company's sector. On the one hand, websites, in some instances, have a predefined list of sector categories, so that companies are required to select one category sector when publishing a vacancy (in some cases more than one category to better describe the company's activities); however, job portals possess their own classification criteria to create a list of sector categories and information between one job portal and another might be not comparable. Moreover, sector categories used by job portals might be highly aggregated. For instance, the Jobportal_b has the category "services". This option is quite broad and very different types of company might be classified under this, and therefore the same, sector.

On the other hand, the job description might contain information regarding the company's sector. However, similar to the above case, companies might use different categories or words to provide information regarding their economic sector. This difference in phrasing categories is an issue as it suggests that the categories or words used by job portals or companies do not adequately describe companies sectors for economic analysis. Fortunately, in most cases, job portals provide alongside the vacancy details the business name of the company that has posted the vacancy. Additionally, in Colombia, the Single Business Registry (RUES, by its Spanish initials) is available[20].

Consequently, it is possible to correlate the vacancy and the RUES database by using company names as a connector between the two databases. However, some challenges present themselves when merging two databases through the use of company names: misspelling or additional information might exist in either, or both, of the vacancy and the RUES database. For instance, in the vacancy database the company's name might appear as "*exito*" while in the RUES database the same company might have been registered as "*éxito S.A*". Thus, both

[20] The RUES is a database where people register their companies so as to pay taxes and receive government benefits. In this database, companies names are available along with other relevant information such as their International Standard Industrial Classification of All Economic Activities code (ISIC).

names in the vacancy and the RUES database might be not the same, even when the databases refer to the same company. This possible difference in names between the vacancy and the RUES database might complicate the merging of the two databases. Given this issue, it is necessary to utilise word-based matching methods (better known as “fuzzy merge” methods) to merge two or more databases based on words or sentences (in this case companies’ names). Generally, word-based matching methods are a set of algorithms that compare sentences and match phrases that are above a certain threshold matching score. The higher the matching threshold, the more accurate the results, but it is possible that fewer observations are matched; the lower the matching threshold, the less precise the results will be, but it is probable that more observations are matched.

Because different approaches exist, each with their own advantages and disadvantages, to identify the economic sector for each job announcement, this thesis implemented a combination of manual coding and word-based matching methods (see Appendix C:).It is important to note that the procedures implemented in this thesis are useful to assign an ISIC code to more than half of observations in the vacancy database (61%). However, the level of disaggregation (4 digits) of this variable might be limited by word-based matching methods or through the use of keywords. For instance, a construction firm might be categorised as “Construction of utility projects” (4220 ISIC code) by observing keyword construction in the company’s name. Although such a company might belong to the civil engineering group (division 42 according to ISIC), at a more disaggregated level, it might belong to the construction of roads and railways (4210) (see Cárdenas 2020b for a more detailed discussion regarding this point).

1.6 Conclusion

Information in job portals has caught the attention of researchers and policymakers insofar as it might help to fill the gap regarding labour demand for skills and, hence, improve skills matching between workers and employers. Nevertheless, to process and analyse information from job portals in a reliable and consistent statistical way is challenging. This section has discussed and proposed different solutions to build a robust vacancy database of job portal information.

Before collecting information from job portals, what is required is a study of the sources to be considered for data analysis. Not every website provides adequate vacancy information. Some job portals provide repeated and/or false information, while other job portals provide a relatively

small number of job announcements. In the case of Colombia, the evidence suggests that vacancy information is well represented in three job portals (Jobportal_a, Jobportal_b and Jobportal_c). It is important to notice that this number can vary from country to country, and over time.

Once the job portal sample is selected, the next challenge is the collection of thousands of job announcements, both systematically and efficiently. The manual collection of information is virtually impossible. Thus, so far, web scraping techniques are the best way to obtain labour vacancy information from job portals. However, to carry out web scraping techniques requires an in-depth understanding of programming (such as R and Python) and the architecture of each job portal selected in the sample. Each website has its unique HTML structure. As a consequence, web scraping techniques involve programming a different algorithm that automatically and periodically collects information from each website. Moreover, websites might change over time. Thus, algorithms need to be updated whenever there is a change in the HTML structure of the websites of interest.

The challenges for analysing vacancy data do not end with the collection process. Job portals provide detailed information regarding job announcements; however, to organise job vacancy information for statistical analysis requires different approaches. Key variables such as wages and required qualifications, among others, are dispersed throughout job advisements. Thus, it is necessary to program an algorithm that deals with linguistic issues (such as gendered words in Spanish), reads each of the job announcements, and creates an indicating variable that takes a value (for example, 1) if a particular pattern emerges on a job advertisement. However, to build a variable such as for a company's sector it is necessary to implement other and more complex text mining techniques such as word-based matching methods (fuzzy merge), and utilise other databases such as the RUES.

Moreover, job portals variables might provide information regarding what occupations (at a detailed level of disaggregation) and skills are demanded at a given point in time. Nevertheless, the implementation of different and more sophisticated techniques and processes is required to deduce and organise skills and occupation information. Thus, the following section will describe the methods that can deduce skill and occupational information, among other relevant variables.

2. Extracting more value from job vacancy information (methodology part 2)

2.1 Introduction

The previous section has shown that job portal information might provide detailed labour demand information such as educational requirements and experience, among other vacancy characteristics. However, what makes job portals a potential and remarkable source of data is that they might provide detailed information in real-time about the skills and occupations demanded by companies. As discussed in Cárdenas 2020a, the dynamic between the skills or occupations offered by individuals, and the skills or occupations demanded by employers is a relevant factor that has strong implications on outcomes for productivity, wages, job satisfaction, turnover rates, unemployment, etc. (Kankaraš et al. 2016; Acemoglu and Autor, 2011). Indeed, the mismatch between the supply and demand for skills might explain a considerable share of unemployment and informality rates in Colombia (see Cárdenas 2020a). Despite the relevance of this topic, detailed information (from official sources such as ONS) for the analysis of the labour demand for skills is relatively scarce due to methodological issues and the high cost of the collection of detailed labour demand information (Cárdenas 2020a). Thus, the key task of this section is to describe the techniques that can be utilised to extract skills and occupational information.

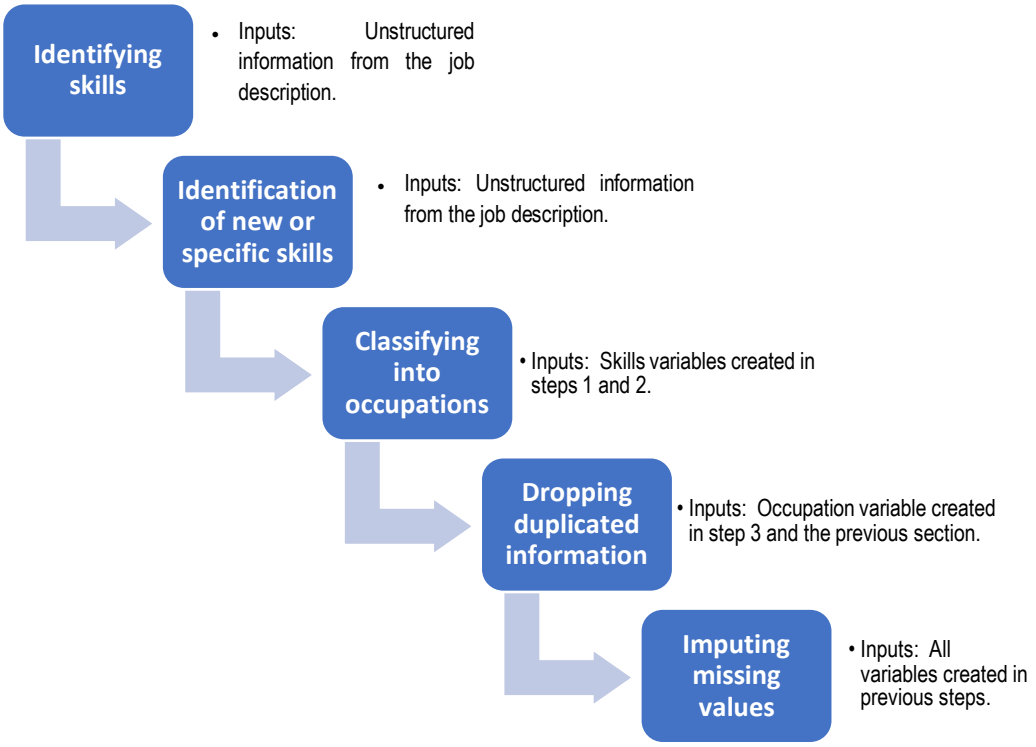
As mentioned in section 1, information from job portals is not categorised with statistical analysis in mind. For instance, non-categorised information related to skills and occupations (for the Colombian case) can be found in job descriptions and the job titles, respectively. Consequently, this section explains the steps required to organise and categorise skills and occupational information from the vacancy database (Figure 2.2 shows the steps that were followed in this thesis to organise Colombian vacancy information). Subsection 2.2 of this section develops a methodology to identify skills patterns in job vacancy descriptions based on international skill descriptors, such as the ESCO (European Skills, Competencies, Qualifications and Occupations). However, there might be some country-specific skills that are not listed in the ESCO's dictionary, or its international skills descriptors might not be updated according to the most current labour demand requirements. Therefore, subsection 2.3 proposes a methodology to automatically identify country-specific or new skills from job portal information.

The classification of job titles into occupations is a critical stage for vacancy analysis. To correctly code the job title variable requires different and advanced data mining techniques. Therefore, subsection 2.4 describes and applies techniques such as manual classification, software classifiers and machine learning to organise job titles into occupational groups. This subsection also proposes a method that uses unstructured information from job titles and skills requirements (variables created in the previous steps) to identify the occupational groups of hard-coding vacancies. With this last procedure, the vacancy database is completely organised.

Once the vacancy database is organised and categorised into occupational groups, educational requirements, etc., it helps to identify duplication problems at this stage. A job vacancy advertisement might be repeated as an employer might advertise the same vacancy many times on the same job portal or between different job portals. Thus, subsection 2.5 deals with duplication issues.

With the vacancy data variables organised and categorised and the duplication problems minimised, as much as possible, an imputation process can be conducted for certain variables. As shown in section 1, vacancy data might contain a considerable number of missing values in the variables of interest (e.g. educational requirements and wages offered). This missing information might create biases in the later analysis of labour demand requirements. Thus, subsection 2.6 outlines how missing values were inputted for the "educational requirement" and "wage offered" variables by using predictors such as occupation, city, and experience requirements, among others (Figure 2.2). Finally, subsection 2.7 presents consolidated, organised, categorised, cleaned and imputed data for the analysis of the Colombian labour demand using job portal sources.

Figure 2.2 Steps for extracting more value from job vacancy information



2.2 Identifying skills

As shown in section 1, in most cases job portals provide abundant information to describe a vacancy. Part of this information is strongly related to the concept of skills: meaning any (measurable) quality that makes a worker more productive in his/her job which can be improved through training and development (Green, 2011) (see Cárdenas 2020a for more discussion on the skill concept). For illustrative purposes, here is an example of a job description[21]:

[21] English translation: "Important agro-industrial company requires a person with basic knowledge in **management systems (ISO, BPM, environmental, SST, RSPO) quality management standards, industrial safety and environmental management, good Excel management and Office automation tools.** Studies: Must have studied **in industrial engineering, administration, microbiology, bacteriology** or be a student in the last year of her/his studies. Experience: minimum of six months in positions or similar experience. Functions: keep updating the **S.G.I of the company, compile and classify, register, distribute and file documentation** which includes physical and electronic correspondence, **write diverse documents** for external internal communication. Salary 836,000 pesos + Social benefits Place of work: Codazzi. interested send resume."

Table 2.4: Job description

Description
"Importante empresa de sector agroindustrial solicita para su equipo de trabajo analista de calidad. La persona debe tener conocimiento básicos **en sistemas de Gestión (ISO, BPM, Ambiental, SST, RSPO) normativa de calidad, Seguridad Industrial y Gestión Ambiental, buen manejo de excel y herramientas ofimáticas**. Estudios: Debe tener estudios en **ingeniera Industrial, Administración, Microbiología, Bacteriología** o estudiante de últimos semestres. Experiencia: mínimo seis meses en cargos o experiencias similares. Funciones: Actualización del **S.G.I de la empresa, recopilar clasificar, registrar, distribuir y archivar la documentación** lo cual incluye correspondencia física y electrónica, redactar documentos diversos para la comunicación interna externa. Salario 836.000 + Prestaciones sociales Lugar de trabajo: Codazzi. interesados enviar hoja de vida actualizada"

Source: Job description taken from Jobportal_a.

As highlighted in Table 2.4, some words or phrases in the job description can be associated with the skills concept. More specifically, words such as "office automation" ("*ofimática*" in Spanish) or "environmental management" ("*gestión ambiental*") can be seen as a precise skill required for this vacancy.

Unlike Lima and Bakhshi (2018) who used pre-defined skills tags to analyse UK job advertisements, for the Colombian case, skills information is not organised under separated variables nor categorised under the same typology. Employers use different words or phrases to describe a skill. Additionally, skills information appears in the job description. Thus, this information needs to be organised to produce informative indicators regarding the labour demand for skills.

As discussed in Cárdenas 2020a, there are different ways (typologies or dictionaries) to organise and analyse information regarding skills. Consequently, the first step to organise the skill information dispersed within vacancy advertisements is to select a dictionary of words or phrases related to skills. Through this method, it is possible to identify the patterns (words or phrases) that are connected to skills in the job advertisements. However, Colombia does not have an official dictionary or a list of skills for such a purpose. Consequently, it is necessary to use international references. In this regard, there are different international skill descriptors available, with, perhaps, the most common skills descriptors being used by O*NET and the ESCO.

As mentioned in Cárdenas 2020a, O*NET is a system based on the US Standard Occupational Classification (SOC) system. This system contains information on hundreds of standardised and occupation-specific descriptors[22]. Importantly, all these job descriptors are available in the Spanish language; thus, O*NET descriptors can be used to identify skill patterns in Colombian job vacancy advertisements.

ESCO is a multilingual classification system, so a Spanish version is available for all European skills, competencies, qualifications and occupations. It is important to note that occupations in the ESCO follow the structure of the International Standard Classification of Occupations (ISCO-08) at the four-digit level, and that the ESCO provides lower levels of desegregation skills for each occupation, such as an exhaustive list of 13,485 relevant skills (skills pillar) (ESCO, 2017). This list of skills might serve to identify those mentioned in Colombian job advertisements.

Moreover, the ESCO list of skills has an important advantage compared to O*NET: since ESCO is mapped following the ISCO-08 structure, the two systems of classification (ESCO and ISCO-08) are compatible. As the ESCO's handbook points out: "This is particularly important because most national occupational classifications are currently mapped to ISCO-08" (ESCO, 2017, p.29). Indeed, in 2015 Colombia accepted recommendations made by the International Labour Organization (ILO) to adopt ISCO-08 as the official classification[23]. Thus, to obtain results compatible with the official national classification for this thesis, the ESCO list of skills was employed to identify skills demanded in Colombian job vacancies.

Once the dictionary was selected, the next step was the implementation of text mining techniques to identify the corresponding skills demanded in job advertisements. Firstly, common words in the Spanish language (such as prepositions, stop words) were removed from the ESCO dictionary and from the job description in the vacancy database. Moreover, all letters were transformed to lower case and words were reduced to their grammatical root in both the ESCO dictionary and in the description of the vacancy database. After this, each word or phrase in the skills dictionary was searched for across each job vacancy advertisement. This exploration of words was encoded into unigram variables (n-gram), which are indicator variables. Variables

[22] See: https://www.onetonline.org/

[23] See: https://www.dane.gov.co/files/sen/nomenclatura/ciuo/RESOLUCION_1518_2015.pdf

take the value of 1 if a certain a word or phrase (pattern) in the skills dictionary is found in an advertisement, and 0 if otherwise.

It is important to notice that each job post does not necessarily contain information regarding skills. There is a considerable share of job vacancies that do not contain skill descriptions. These missing values do not mean that an employer does not require any skills for a particular job, as employers always need workers with a set of skills. However, when publishing a vacancy, employers might not consider it necessary to explicitly write a list of the skills required. Consequently, as will be discussed in more detail in Cárdenas (2020b), unigram variables show the key skills needed for a vacancy, but they do not sufficiently identify the complete set of skills needed to perform a job.

Thus, the identification of skills mentioned in the job description helps to identify the key skills in demand within the Colombian labour market. Additionally, as shown in subsection 2.3, unigram skill variables will serve to identify new or specific skills that are requested in the Colombian labour market, and that are not listed in the ESCO dictionary of skills. To have a complete identification of required skills it is necessary to classify job titles according to an occupational classification (see subsection 2.4). Moreover, as will be seen in subsection 2.4.7, unigram skill variables facilitate assigning occupational codes to the vacancy database.

2.3 Identification of new or specific skills

Although the ESCO dictionary of skills is a complete list for the European labour market, there might be some country-specific skills which are not listed. For instance, Colombian employers might demand different skills compared to Europe. This issue might be the case regarding a specific technology (e.g. software) that is demanded in Colombia, but not used in Europe. Moreover, as mentioned in Cárdenas (2020a), updating dictionaries or occupational classifications might require substantial time, while labour markets rapidly change. This time lapse between changes in the labour demand for skills and the time needed to upgrade skill dictionaries might cause those skills dictionaries to not adequately measure what current skills are in demand.

Consequently, to identify new skill patterns from job descriptions it is necessary to discard information that does not refer to any skill. As in the previous section, common words in the

Spanish language (e.g. stop words) were removed from job descriptions. The above technique diminishes a considerable number of words not related to skills; however, a significant number of words might remain that are not relevant to the identification of new skill patterns. As a consequence, a stop words dictionary was created for this study based on the information available in Colombian job vacancies to continue removing non skill-related words. More specifically, column variables from the vacancy database, such as city, wages, type of contract, among others (not related to skills), were used to build a stop words dictionary. The words that appeared in this new stop words dictionary were removed from the description of each vacancy. Nevertheless, several words might remain that do not correspond to new skill patterns. For instance, skills identified with the ESCO dictionary remained in the description of the vacancy; consequently, the ESCO skills dictionary was used as a stop words dictionary to remove those skills that were identified previously in subsection 2.2. Hence, the words that remain in the description of the vacancy might provide relevant information regarding new and/or specific skills demanded by the Colombian labour market.

2.4 Classifying the vacancies into occupations

One of the most critical variables is "job title" because it summarises the main characteristics of the demand for labour and allows classification of the jobs into occupations (or skills). For example, if a vacancy requires an "Accountant" job title, then people who are interested in applying for that vacancy need to know accounting, mathematics, how to process information, amongst others. Over time, if there are not enough people to fill this job title ("hard to fill vacancies"), then the government could take action to inform individuals with these skills or prepare people for those occupations (or skills). But first the appropriate government department needs to be informed of any unfulfilled job skills requirements in the labour market.

According to Figure 2.3, in Colombia, January 2017, the most frequent words that appear in the "job title" variable, and, as a consequence, the most demanded jobs for that time period were: assistants ("*auxilar"*), salespeople ("*venta"*), engineer ("*ingeniero"*), call centre employees *("call center"*), customer service ("*cliente"*), manager ("*supervisor"*), drivers ("*conductor"*), among others.

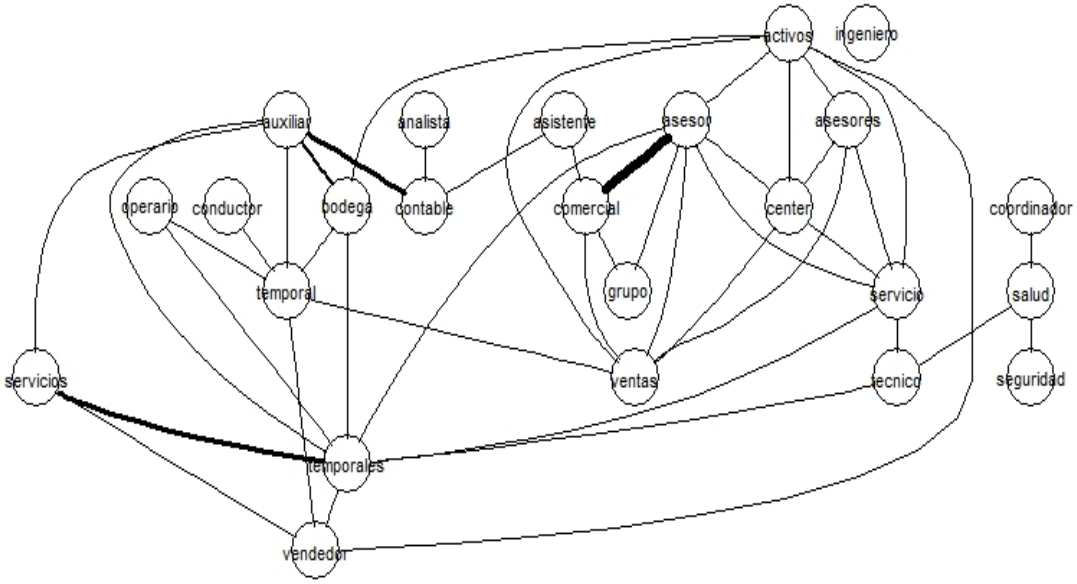
Figure 2.3: Word cloud: Frequency analysis[24]

Source: Vacancy data base 2017. Own calculations.

Figure 2.4 shows the words that are most associated with job titles in more detail. A word is related to another word if both words frequently appear together in a job title. Consequently, the thicker the black line in Figure 2.4, the stronger the association is between words. For instance, within the group "assistants" ("*auxiliar*"), the most common job title is "accountant assistants", followed by "warehouse and services assistants". Within the "advisor" ("*asesor*") group, the most frequent occupations are in "sales, commerce, and customer service".

[24] The text mining figures are presented in the Spanish language because it is the original language used in the Colombian job portals.

Figure 2.4: Word association: Frequency analysis



Source: Vacancy data base 2017. Own calculations.

The above figures are an approach to distinguish the most commonly demanded job titles. However, these figures have many limitations. For instance, they do not identify synonyms. As shown in Figure 2.3 and Figure 2.4 "*assistant*" and "*auxiliar*" are considered as different categories, though they can, on many occasions, refer to the same job category. Moreover, they also count plural and singular titles as different categories (e.g. Assistant and Assistants). To avoid these mistakes and for statistical purposes, it is necessary to use an occupational classification which is defined as a "tool for organising jobs into a clearly defined set of groups according to the tasks and duties undertaken in the job" (ILO, 2017).

Regarding job title, this research seeks to classify all the information available to the ISCO-08[25]. However, as Štefánik (2012) points out, there are challenges in transforming job titles into occupation categories because they were created for other purposes. In some cases, there is going to be more or less information required to classify job titles into occupations. However, such challenges are present in all types of sources such as household or company surveys that

[25] As previously mentioned, Colombia accepted the recommendations made by the ILO to adopt ISCO-08 (ILO, 2008) as an official classification for jobs.

collect information on occupational titles. Nevertheless, in the case of vacancy data collected from the Internet, to classify job titles into occupations might be more difficult. For instance, alongside job titles might appear the company's name, the city where the vacancy is available, among various words that are not directly related to job title information. Moreover, as mentioned above, companies might use a variate of different words to describe the same occupation. This issue makes the classification of job titles into occupational codes a challenge. However, to correctly assign occupational codes is the most important quality of a vacancy database to be able to utilise it for statistical purposes.

Given the complexity of classifying job titles into occupations and the importance of this information for the researchers, government and other institutions, the economic and statistic literature has used three tools to perform the classification process: manual classification, classifiers (Cascot or O*NET API), and machine learning. Manual classification refers to the process where a person or group of people observe job titles. Traditionally, Gweon et al. (2017) remarks, assigning occupational codes to texts (job titles) has been a manual task performed by human coders. However, manual classification is a time-consuming and expensive process, especially when handling large databases such as Colombian vacancy data[26].

Additionally, to guarantee a certain level of coding quality, this manual process would require a professional knowledge regarding occupational classifications and occupation titles. Nevertheless, as Gweon et al. (2017) highlight, manual classifications might provide inconsistent results even with the use of professional coders. Consequently, coding accuracy depends on people's knowledge, occupational criteria (that might change between people), and the words used to describe a particular job.

More recently, the use of partially or completely automated coding has arisen. Both partial and complete automatic coding significantly reduce coding time. The former term refers to a process where researchers use software to set different rules in order to classify certain occupations. For instance, if words such as clerk-bookkeeper or assistant accounts appear in the job title or job description the set of rules would classify those job titles as "Accounting and bookkeeping clerks"

[26] For instance, the Colombian vacancy data collected for this thesis in November 2017 consists of around 28,820 job titles (after dropping duplicated titles), and the manual classification of these titles would require a considerable amount of time for a person or a group of people.

(using ISCO-08). The latter term, completely automated coding refers to methods such as machine learning. Briefly, these set of techniques work in the following way: there is an initial stage where the algorithm requires a (representative) training database in which a set of job titles exist which are already properly classified into occupations (perhaps manually classified). Based on this database, the algorithm "learns" rules of association to code job titles. With this knowledge, the algorithm can predict the most probable occupational code for each job title for new data (Gweon et al. 2017; Lima and Bakhshi, 2018).

Moreover, software such as Cascot (Computer Assisted Structured Coding Tool[27]) exists (Jones and Elias, 2004) (see Subsection 2.4.3), that allows both partial and/or complete automatisation. This kind of software already contains a set of logic rules. Based on a score of similarity between occupation titles (provided by the occupational classification, e.g. ISCO-08) and job titles (e.g. posted on job portals) the software assigns a corresponding occupational code (which has the highest similarity score). In this way, a list of job titles can be automatically classified. However, complete coding automatisation was still a challenging process at the time when this thesis was written due to the complexity of categorising occupational titles (Gweon et al. 2017). Besides, algorithms fail to provide a perfect classification for each job title (Belloni et al. 2014)[28].

Thus, given the availability of several tools to classify occupations and the advantages and disadvantages of each one of them, I will now discuss manual coding, cleaning, automatisation, and adapting Cascot.

2.4.1 Manual coding

As pointed out before, manual coding is a time-consuming task. However, as shown in Figure 2.3 and Figure 2.4, there are some job titles which are more frequently mentioned by employers, hence those job positions constitute a significant share of the vacancy database. Additionally,

[27] Developed at the University of Warwick by the Institute for Employment Research.

[28] To avoid misclassifications, Jones and Elias (2004) recommend the implementation of both partial and fully automated coding (semi-automatic coding). For instance, in the Cascot case, the authors suggest automatically classifying all job titles (inputs) and keeping a record of similarity scores. For those job titles where the similarity score is below a minimum threshold, it is necessary to assign a corresponding occupational code manually. In this way, the time spent classifying job titles into occupations will decrease, and a certain level of coding quality will be guaranteed.

automatic algorithms might misclassify some job titles as automatic methods of classification might fail in classifying some job titles that appear with more frequency in the vacancy database. As a consequence, coding quality might be primarily affected by the misclassification of some common job titles.

In order to ensure that the most frequent job titles are adequately classified, a careful and manual coding process was carried out for job positions which were more numerous and, therefore, it was relatively easy to determine their occupational group. Moreover, as words in the Spanish language are gendered and words might slightly differ in the plural and the singular, the roots (patterns) of the words were used instead of looking for exact combinations of words. For instance, manually classified titles such as "accountants" were extracted by using the root "*Contador*" instead of "*Contadora*" for a woman or "*Contadores*" in the plural case. By doing so, a total number of 50 job titles received an occupational code (which corresponds to around 27% of the job advertisements). This information suggests that a considerable share of Colombian vacancy information is concentrated across relatively few job titles.[29]

2.4.2 Cleaning

As mentioned above, coding quality depends on the tool used and on the quality of the input data. However, job titles displayed in job portals might contain extra information (noise) that might affect coding quality. While there are some group words such as prepositions that might be easy to identify and clean from the data, there are other words that do not belong to a specific group of words that frequently appear in job titles and do not describe a job position.

As shown in Figure 2.3, in the job titles, abundant information is not directly related to the job position (such as company name and working hours). It is common to see, words such as "time", "immediately", and "required", among others, in the Colombian vacancy data. The presence of these words might affect the performance of automatic classifiers. To assign an occupational code, tools, such as Cascot or the ONS Occupation Coding Tool, compare the similarity of words in the job title from a job vacancy (or another source of information) with a directory of job titles. The extra information might affect this comparison. For instance, when the input is "accountants"

[29]At this point, this result neither validates nor invalidates the reliability of the data. The Colombian labour market might demand a particular set of occupations (see Cárdenas 2020b and for further discussion).

with a similarity index of 92 Cascot assign the ISCO code 2411 (“Accountants”)—in a scale of 0 to 100, the higher the number, the higher degree of certainty that a given code is the correct one. However, when the input is “accountants immediately” the similarity index drops to 66.

Moreover, machine learning models might also be affected by extra information provided by the employer. Usually, for carrying out a machine learning model, a robust, cleaned and (manually) classified database is required. This database is divided into two: the training database and the test database. The “training database” works as a reference guide, and the computer learns how to classify job titles into occupations. The “test database” checks the algorithm’s performance; in that, the algorithm is applied to the test database using inputs such as the job title to predict the occupational code. Subsequently, the result of this automatic classification is compared with the occupational codes that were previously assigned in the database. The comparison is summarised in a “confusion matrix” which shows the number (or percentages) of matches between automatic classification and the number that was previously in the test database. The higher the number of matches, the better the performance of the algorithm.

Nevertheless, when a new database requires to be classified into occupational codes and job titles contain noisy information (supposing that the job title is the only input for the machine learning algorithm) the performance of the algorithm might decrease. The algorithm learns classification rules based on a training database that might not contain noisy information. Consequently, when new observations contain extra information the algorithm might not correctly understand that the additional information does not provide relevant information to assign an occupational code to, and the probabilities of misclassification might increase.

Thus, before conducting automatic classification processes, the job title variable, which is the primary input to assign an occupational code, was carefully cleaned. First, prepositions, adverbs, nouns, among others, were dropped from the data. Second, the variables “city” and “companies’ name” (provided the structure of the website contained this information) were used to identify all possible locations and employers’ names that might arise in the job title variables. With this process, names were dropped that might appear in the job title. Third, with a visual inspection of the vacancy database and the usage of word clouds, it was possible to identify and drop those words that did not contain information regarding occupation in the job title. After this manual and cleaning process, I proceeded to use automatic classification tools and techniques.

2.4.3 Cascot

The first step in the automatisation process is the usage of Cascot. As mentioned before, this tool was developed by Jones and Elias (2004) at IER. Cascot is designed to assign an (occupational or industrial) code to texts. In the case of occupational classification, Cascot allows the classification of a piece of text (job titles) according to their UK Standard Occupational Classification (SOC 1990; 2000; 2010). Moreover, since 2014, a multilingual ISCO-08 version of this computer program has been developed for nine languages (Dutch, English, Finnish, French, German, Italian, Portuguese, Slovak and Spanish). Additionally, in 2016, the software was extended to another five languages (Arabic, Chinese, Hindi, Indonesian and Russian).

This multilingual capability is one of the most critical characteristics of Cascot. It allows classifying job titles from different languages into occupations following an international standard such as ISCO-08. In order to classify a piece of text into an occupational classification (e.g. ISCO-08), Cascot has a set of rules—such as downgraded words, equivalent word ends, abbreviations, replacement words, word alternatives, etc. (IER, 2018)—which reveal the best matches between job titles (inputs) and occupational classifications with corresponding similarity scores. Importantly, to set up all the association rules (mentioned above), the IER made partnership arrangements with experts for each country covered for the testing and refining of Cascot (Wageindicator, 2009).

Moreover, Cascot outputs have been compared with high-quality and manually coded data (Jones and Elias, 2004). According to this test, 80% of records that receive a similarity score higher than 40 coincided with the manually coded data. Thus, Cascot offers, to a certain extent, a well-defined directory of job titles with occupational codes and association rules that can be used for coding job titles.

Consequently, one of the main reasons to use Cascot is that it already has a depth and reliable knowledge base, built over years. Indeed, relatively new classification methods such as machine learning should consider and "learn" from the association rules that have been created through years of research using Cascot. Moreover, this tool has a considerable advantage in a context where there is not (or at least not publically available) a trustworthy pre-processed database with job titles and occupational codes. Machine learning methods need as an input a training

database (which is a data that was previously and correctly classified). Without this training database it is not possible to use machine learning models to assign occupation codes.

Taking the above reasons into account, Cascot was used to classify job titles in the Colombian vacancy database. Following the recommendations of Jones and Elias (2004), Cascot assigned an occupational code to a job title if the similarity score was greater than 45. This threshold was to re-ensure that the Cascot outputs would coincide with the manual coding revision in most cases. By doing so, around 38% of observations in the vacancy database received an occupational code at the four-digit level[30]. Thus, 35% of job advertisements required further data management to assign a proper occupational code.

2.4.4 Revisiting manual coding (again)

Provided that 35% of the database was "hard-coded" (not classified by Cascot), it was necessary to conduct another short manual-coding process. Here, the same methodology was applied that was explained in Subsection 2.4.1 First, a visual inspection of the vacancy database was conducted on the data that was not classified by Cascot. Job titles that appeared more frequently in the database were manually assigned an occupational code. Once again, the usage of the roots of the words was necessary to avoid any gendered or plural (singular) issues. With this, it was ensured that hard-coded job titles that were more frequent in the vacancy database received a proper occupational code. In total, 50 job titles were manually coded, which corresponds to around 5% of the total number of job advertisements. At this point, approximately 70% of observations were assigned an occupational code with a relatively high standard level of confidence.

2.4.5 Cascot adaptation according to Colombian occupational titles

The ISCO contains a standard list of occupational titles used in the international workplace which is linked to categories in its classification structure. This list is a key input for Cascot to match occupational codes and job titles. However, as mentioned by ILO (2008, p.68): "[occupational titles provided by ILO] might be a good starting point to develop a national index. The national

[30] A sample of those observations was selected to evaluate the accuracy of the Cascot tool for the Colombian case. According to this manual check, around 94% of observations had the correct occupational code (ISCO-08) at a four-digit level. Moreover, common mistakes were manually corrected.

index, however, needs to reflect language as used in survey responses in the country concerned". Even in counties with the same language, job positions might be named differently depending on the national context[31]. Consequently, standard occupational titles provided by ILO might not cover a considerable share of Colombian job titles, hence Cascot might not assign an occupational code to a high portion of Colombian job titles. Indeed, this issue of context might explain that at this point, only 38% of job portal observations were categorised using Cascot.

Moreover, DANE released an adaptation of the ISCO occupational titles according to the Colombian context in 2015 (DANE, 2015). Additionally, Cascot can be edited and, hence, the adjustment of the Colombian occupational titles can complement this tool. Consequently, the following step was updating Cascot to the Colombian context by using the occupational titles utilised in this country. Once this adaptation was made, the job titles that were not coded in the previous steps (around 30% of the total number of job advertisements) were processed once again for Cascot with the same specifications mentioned in Subsection 2.4.3. Interesting, with this adaptation of the tool, around 12% of the total number of advertisements were assigned an occupational code. Thus, by only adapting the Cascot tool with national occupational titles of Colombia, the portion of job advertisements considerably increases from 70% to 82%.

However, concerns might arise regarding the accuracy of coding with this adapted version of Cascot. Regarding this concern, it is necessary to highlight that the occupational job titles used to adapt Cascot come from the national statistical department in Colombia and are publicly available. Moreover, the list of Colombian job titles is the product of joint work by institutions such as DANE, the Ministry of Education, the Ministry of Labour, and training providers, among others (DANE, 2015). Thus, the input "occupational titles" should be similar to job titles in job advertisements[32].

[31] For instance, in Colombia, there is a particular job title to define general maintenance and repair workers, which is "todero". This job title cannot be found in countries such as Perú or Chile (where Spanish is also the official and most spoken language).

[32] A manual check was carried out to determine the accuracy of correctly coded observations. According to this manual check, around 92% of observations had the correct occupational code (ISCO) at a four-digit level. Moreover, common mistakes were manually corrected.

2.4.6 The English version of Cascot

As a result of the above manual check, a considerable portion of job titles that were found to lack an occupational code were those written in English. Despite Spanish being the official language of Colombia (among other minority indigenous languages), job titles such as "customer care analyst", "data analyst", "courier", etc., are written in English. Consequently, the English version of Cascot might help to classify some of the job titles in the vacancy database. However, the English version of Cascot assigned an occupational code to a job title if the similarity score was greater than 60. This threshold is set at 60 to avoid any confusion and misclassification with job titles in the Spanish and English Cascot version. By doing this, 3% of job titles in the vacancy database received an occupational code.

At this point, 15% of observations remained without an occupational code. There were three options for classifying the remaining job titles: 1) manual coding, 2) using lower minimum similarity threshold through Cascot, or, 3) other techniques such as machine learning. The first method, as mentioned more than once above, is a time-consuming task. Therefore, this option was not considered. Meanwhile, the second and third options contain various advantages and disadvantages. On one side, the Cascot similarity threshold could be lowered to classify more job titles (so far, the threshold used has been 45). Nevertheless, this might increase the number of misclassified observations[33]. On the other hand, machine learning techniques could serve as a complement to identify occupations. As mentioned previously, machine learning techniques have been implemented during the last year to assign occupational codes to job titles. Depending on the sophistication of their algorithms and inputs (training and test databases), this technique might adequately assign occupational codes to job titles (Bethmann et al. 2014).

[33] This option is the most straightforward alternative to assigning occupational codes to the remaining observations because it is relatively easy to conduct. Although Jones and Elias (2004) recommend using a minimum threshold of 40, each researcher can reduce this threshold and increase the number of observations with occupational codes. However, this might also increase the number of misclassified observations. The Cascot minimum score threshold was lowered to 30. This minimum threshold was set arbitrarily as a starting point to evaluate Cascot's performance. A sample of observations with a threshold of 30 was taken to assess Cascot's performance. As expected, the accuracy level of automatic coding decreased. Around 39% of job titles were incorrectly classified. Thus, lowering the Cascot threshold was not an option to classify the remaining job titles.

However, for the Colombian case, machine learning algorithms might fail in assigning occupational codes. Ideally, any training dataset should contain a representative sample of job titles which are correctly coded; unfortunately, in Colombia, there is not a sample of vacancy job titles with occupational codes. The closest sample is the Colombian occupational job titles mentioned above. This input might serve as a training database. Nevertheless, as shown when the adapted Cascot was used, Colombian occupational titles might not be a fully representative sample of job titles: otherwise, the number of observations with occupational codes assigned by Cascot would have been higher.

Consequently, the machine learning algorithm might not assign an occupational code to those job titles which are not represented in the training database. Indeed, if the adaptation of Cascot did not appoint an occupational code to 15% of the vacancy database, machine learning algorithms might provide a similar result. Machine learning algorithms face the same problem: in Colombia, there is not a database that serves as an example to classify challenging job titles.

2.4.7 Machine learning

The use of machine models that classify job titles into occupation codes has arisen over the last decades. As Gweon (2017) highlights, institutions such as the Australian Bureau of Statistics have favoured this method. In concrete terms, machine learning is a "set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" (Murphy, 2012, p.1)[34]. Moreover, as Murphy (2012) points out, classification (Supervised Learning)[35] is perhaps the most commonly used form of machine learning to solve real-world issues. The idea in this method is to classify a "document" for instance, a job title, (usually denoted as D) into one of several classes (C) based on some previously learnt training inputs (X). The computer determines how to classify a document based on both a training dataset and a particular association algorithm. The former refers to a pre-processed dataset with an N number of training examples (usually denoted by

[34] These machine learning methods have been applied in several fields, such as health and economics, among others (Varian, 2014; Zhang and Ma, 2012).

[35] Unsupervised and reinforcement learning are other types of machine learning algorithms. However, as the purpose of this subsection is classification of job titles, this thesis is focused on Supervised Learning.

D). For the job titles case, this database is a pre-processed database with job titles assigned with corresponding occupational codes (see Appendix D:).

In terms of assigning occupational codes to job titles, the economic and statistic literature has favoured SVM (Supporting Vector Machines) (Gweon et al. 2017) (see Appendix E:). However, as mentioned in the Appendix F:, 40% of the vacancy job titles were incorrectly classified by using SVM. Therefore, the SVM machine learning algorithm which only uses job titles is not an option to classify the remaining observations in the vacancy database.

2.4.7.1 Nearest neighbour algorithm using job titles

As shown in the previous subsection, the numeric transformation of job titles with the SVM algorithm might serve to assign an occupational code to hard-coding observations. However, the number of job titles classified by the SVM is limited, and, consequently, it is necessary to use more advanced techniques to code job titles. In this regard, Gweon et al. (2017) demonstrate that (with some adaptations) the nearest neighbour algorithm might provide better results regarding accuracy than the SVM algorithm. Briefly, the nearest neighbour algorithm takes new record(s) (in this case n-gram of a job title), maps this (these) new record(s) in the training dataset, and finds the closest observation to this new record based on n-grams of the job titles. Once the nearest neighbour(s) is (are) selected, the algorithm assigns to the new record(s) the class (y) of its (their) closest neighbour(s) (see Appendix G:).

2.4.7.2 Machine learning using skills

Conversely, Lima and Bakhshi (2018) proposed an extension of the basic machine learning model for classifying job titles into occupations. The authors used UK job vacancies published in 2015, collected by Burning Glass[36]. This company assigns each vacancy one or more of 9996 tags derived directly from the job advertisement text (the authors did not clarify how and based on what the tags were built). Consequently, instead of using job titles (n-grams) as an input to assign an occupational code to each observation, the authors propose to use a naïve Bayes algorithm that takes as its predictors (x) the skills mentioned in the vacancy advertisement. By

[36] Burning Glass is company that provides job market analytics.

doing so, Lima and Bakhshi (2018) demonstrated that a skills-based classifier might improve the coding of jobs titles that are poorly classified.

2.4.7.3 Nearest neighbour algorithm using skills and job titles

Provided the above advantages and limitations of the more recently proposed algorithms, this thesis uses an extension of Gweon et al.'s (2017) algorithm by adding the n-grams (input x) information related to skills as suggested by Lima and Bakhshi (2018). Specifically, it is recommended to complement n-grams (input x) from the job title with the skills mentioned in the job description. Skills information is supposed to be highly correlated with job title. For instance, for a job position such as "Secretary", it is logical to think that employers will require relatively more skills related to office automation, while for a job position such as "Kitchen helpers" the skill requirements will be relatively more related to food production. Consequently, by considering the skills demanded and the job titles, it is possible to find a more similar training dataset that might improve automatic coding (see Appendix G:).

2.4.7.3.1 Application of the extended-nearest neighbour algorithm to the vacancy database

As mentioned in subsection 2.4.6, 15% of job titles remained uncoded at this stage by manual and Cascot procedures. Consequently, the final step to classify the remaining job titles was conducted the extended-nearest neighbour algorithm explained in Appendix G: (Tables G.5 and G.6). However, as pointed out in subsection 2.2 of this section, unlike Lima and Bakhshi (2018) where the authors had at their disposal pre-defined skill tags to use as inputs for the machine learning model, for the Colombian case skills information (which is the key input required to implement an extension of the nearest neighbour algorithm) is not organised into separate variables, nor categorised under the same typology. Thus, this thesis uses the n-gram skill variables created in subsection 2.2 as an input for the algorithm proposed here.

By doing so, this thesis uses an algorithm (nearest neighbour) with a proved high accuracy level for categorising job titles. Moreover, using skill n-grams based on the ESCO dictionary shows that the description might increase the accuracy level and the number of job titles coded without the need for pre-defined skill tags (Appendix G:). With this method, 10% of job titles were coded.

Consequently, at this point, 95% of the job titles in the vacancy database have received an occupational code.

Despite machine learning methods and classifiers such as Cascot significantly reducing the time spent on coding, at the time this thesis was written it is still necessary to use manual coding for those job titles which remain unclassified. Consequently, 50 job titles were coded manually. Thus, through automatic and manual processes, 96% of the job titles were coded according to ISCO (4-digit level)[37].

2.5 Deduplication

Along with the categorisation challenges shown above, there is another important issue to consider, which is the possibility of duplicated information. As the data is collected from different websites, some job advertisements can appear on more than one job board, or even on the same job board (Cárdenas 2020a). This issue can result in a significant over-counting of job advertisements and might affect the results when the data is analysed. For those reasons, before data analysis it is necessary to apply a measure to identify which vacancies are duplicated to discard all but one of them. This process is known as "deduplication." (Carnevale et al. 2014).

One option is to drop those vacancies which have the same job title, level of education, city, sector, date published, wages, etc. However, this string-based approach is not enough to completely solve the duplication problem, e.g., an employer can post a vacancy with the job title “Taxi Driver” on a website, and another website can write “Taxis Driver” for the same vacancy. With the method described above, this vacancy would count as a different one. Therefore, it is necessary to develop or adopt a measure of “similarity” to decide the probability with which an observation is duplicated. In this regard, Gweon et al. (2017) have shown that n-gram-based

[37] Importantly, a significant percentage of non-classification might be explained by the absence of key information in the job title variable. The most frequent words in those job titles without an occupational code do not provide information regarding the job position. For instance, a regular word is “bachilleres” (which in English means “undergraduate”). Clearly, with only these kind of words in the job title it is not possible to identify their requirements through automatic or manual means. One reasonable alternative to overcome this issue is to take into account the job description. Perhaps, information about the job position is in the description rather than the job title. Thus, processing and identifying specific patterns in job descriptions might increase the number of observations with an occupational code. This further development will be a part of future work.

methods for dropping duplication in job titles are preferable than string-based methods. As mentioned in subsection 2.2, n-grams are a set of indicator variables based on text patterns. The variables take the value of 1 if there are specific patterns.

Consequently, ngram-based methods are not sensitive to minor changes in string variables (such as the job title). Thus, following Gweon et al. (2017), an n-gram based method was applied to drop the maximum number of observations duplicated. More specifically, a duplicated job advertisement was discarded if the values of dummy variables previously created (such as experience, educational requirements, type of contract, localisation and wages) were the same as other job advertisements, including their ISIC (section 1) and ISCO codes (subsection 2.4 of this section), the publication date and the number of job positions required. By doing so, around 10% of observations were discarded.

2.6 Imputing missing values

Provided that the information comes from websites and employers who might not provide a full description of the vacancy, variables exist with missing values. For instance, despite the text mining techniques explored in section 1, around 30% of observations in the "wage" variable have missing values. As the presence of missing values can create biases in the analysis (Little and Rubin, 2014), it is essential to implement imputation techniques to analyse the full data vacancy information.

In this regard, Carnevale et al. (2014) with hot-deck and cold-deck methods imputed missing education requirements in job advertisement data using a combination of the education distribution of the vacancy (no missing values) data, and the education distribution of employment (from the American Community Survey—ACS). With such a method, they demonstrated that it is possible to use the whole vacancy database to test if the information contained in it is representative of different education levels.

Given the relative importance of the analysis of labour demand for skills and the considerable presence of missing values in the data, for this thesis an imputation procedure is conducted for the wage and educational variables.

2.6.1 Imputing educational requirements

For the Colombian case, 20% of observations in the educational requirement variable contain missing values. These missing values do not mean that for those vacancies Colombian employers do not have any educational requirements. Employers might forget to mention educational requirements or information regarding education might be implicit in other variables (such as the job title). Indeed, in most of the job titles in the vacancy database the educational requirements are implicit. For instance, job titles, such as lawyer, economist, and psychologist, among others, implicitly reveal that employers require a worker with at least university education.

Consequently, to impute the missing values a hot-deck imputation was conducted as proposed by Carnevale et al. (2014). Specifically, through this method an observation with a missing value in a particular variable receives a value which is randomly selected from a sample ("deck") of non-missing records that have some characteristics ("deck variables") in common with the observation with the missing value. For instance, for the Colombian case an observation with a missing value in “educational requirement” receives a value from an observation which is randomly selected from a sample of records that have the same characteristics in common, such as the same occupation. Consequently, as a first step, it is necessary to define which characteristics define the sample of donors ("deck") for an observation with a missing value.

Within a vacancy, this variable’s occupation, city and year were considered as characteristics which defined the sample of donors. By using these three variables, it is possible to establish a proper sample of donors for observations with missing values for their educational requirements. The occupational variable (at a 4-digit level) guarantees that both the donors and the missing observation(s) contain similar skills and tasks. Indeed, the occupational variable is the most import factor of the imputation process because, as mentioned above, occupation (job title) is a concept strongly related to educational requirements.

Additionally, examining the city (where the vacancies were posted) controls possible differences in educational requirements from one place to another (e.g. a city to a town). The year of the vacancy controls for the fact that educational requirements change over time. As Spitz-Oener (2016) notes, to perform a particular occupation today involves greater complexities than at the end of the 1970s. For instance, in the past, it was enough to have a high school certificate to apply for a job as a secretary, now for the same job title is necessary to have a higher educational

level given technological changes, among other factors. Moreover, no other characteristics in the vacancy database were taken into account due to the high presence of missing values in those variables (e.g. wages).

Thus, an observation with a missing value in "educational requirement" receives a value from another observation if, and only if, that record was offered in the same city and year and has the same occupational category. It is important to note that this thesis did not implemented the cold-deck method. In contrast with the hot-deck method, cold-deck imputation picks donors from another database; for instance, from household surveys. This thesis do not to use the cold-deck method for the following reasons. First, the frequency of missing values in educational requirements is not as high compared to the study by Carnevale et al. (2014) where roughly 50% of the vacancies have a missing value in their educational requirements. Thus, for the Colombian case, there is enough information with no missing value (80%) to impute the remaining missing values.

Second, and more importantly, the cold-deck method proposed in Carnevale et al. (2014) uses the American Community Survey (ACS) (which is a labour supply survey) to impute missing values in the job vacancy data. However, as will be discussed in more detail in Cárdenas 2020b, missing vacancy values based on a household (supply) survey might be problematic due to the distribution of educational requirements (among other characteristics) that might differ between labour demand and labour supply. Moreover, part of this thesis seeks to test if the vacancy database shows consistent patterns compared with official statistics such as household surveys. Consequently, the implementation of a cold-deck method with a household survey imposes, on the vacancy database, a distribution of educational requirements related to labour supply, and thus any comparison in terms of educational level between labour demand and supply might be affected by the cold-deck imputation process.

2.6.2 Imputing wage variable

Finally, given the importance of wages for labour demand analysis and the presence of a missing value for this variable in the Colombian vacancy database (around 30% of total observations), an imputation procedure was conducted. Traditionally, imputation methods involve linear or logistic regressions; however, as Varian (2014) mentions, when a large amount of data are available, better methods to impute variables such as the LASSO regression ("least absolute

shrinkage and selection operator") can be applied. Unlike linear models, the LASSO model penalises the predictors that do not have relevant information and might increase the error term (e) for predicting an output (y)—in this case the missing values for the wage variable (Varian, 2014). In other words, the LASSO model selects and drops those predictors (variables) that do not contribute to wage prediction.

The occupation variable might be comprised of 40 different values (sub-major ISCO groups), for instance, which means that for the LASSO model those values in the occupation variable are transformed into 40 dummy variables. Specifically, to impute the wage variable (y) in the vacancy database the following was conducted:

$$y = \beta_i \ Occupation_i \ \chi_{\{i=1\dots40\}} + \beta_i \ county_i \ \chi_{\{i=1\dots32\}} + \beta_i \ quarter_i \ \chi_{\{i=1\dots4\}} + \beta_i \ education_i \ \chi_{\{i=1\dots8\}} + \beta_i \ Workday \ \chi_{\{i=1\dots3\}} + \beta_i \ TypeContract_i \ \chi_{\{i=1\dots4\}} + \varepsilon$$

Where y is the wage variable, "$occupation$" denotes the set of dummy variables which identify occupation (ISCO—two-digit level, 33 subgroups)[38]; "$county$" represents the set of dummy variables which identify the county where the vacancy is available (there are 32 counties in Colombia); "$quarter$" denotes dummy variables that indicate the quarter of the year when the vacancy was downloaded; "$education$" represents a set of dummy variables which indicate educational requirements (six categories[39], see Table 2.5); "$Workday$" and "$TypeContract$" are sets of dummies variables indicating the workday (three categories, see Table 2.5) and the type of contract (four categories, see Table 2.5) offered by employers[40].

[38] The occupation variable was grouped at a two-digit level to avoid oversaturation and due to computational limitations.

[39] Due to frequency issues, specialisation, master and doctor's degree categories were grouped in one category: "postgraduate".

[40] The variable sector was not included in the imputation model due to the high frequency of missing data.

2.7 Vacancy data structure

The above sections, along with section 1, provide a robust methodology to process and organise job portal information. As a result, the Colombian vacancy database created in this thesis has the following structure:

Table 2.5: Basic data structure

Variable	Definition
Job title	Short description about the job title offered
Vacancy description	Detailed information about the profile required to fill the vacancy
Labour experience	Dummy variable, it takes values of 1 if the vacancy (explicitly) requires any labour experience and 0 otherwise
Number of vacancies	Number of job positions offered for each job advertisement
Company name	Name of the company who published the job advertisement
Publication date	Starting date when the job advertisement was placed
Expiration date	Date when the job advertisement expires
Educational requirements	Set of dummy variables that identify the educational attainment required to fill the vacancy: a. primary; b. bachelor; c. lower vocational education; d. upper vocational education; e. undergraduate; f. specialisation; g. master; h. doctor's degree. See Cárdenas 2020c.
Wage	Continuous variable which indicates the amount of money that the hired person will receive
Imputed Wage	Continuous variable which indicates the amount of money (imputed) that the hired person will receive
Type of contract	Set of dummy variables that identify the type of contract offered by the employer: a. fixed-term contract; b. indefinite duration contract; c. freelance; d. by activities
Workday	Set of dummy variables that identifies the workday offered by the employer: a. full-time; b. part-time; c. by hours
City	Place where the vacancy is available
Sector ISIC	ISIC Code (2 digits if possible)
Skills	Set of dummy variables that identify the skills required by employers according to ESCO
Specific skills	Set of dummy variables that identify (country-specific) skills required by employers and are not listed in the ESCO dictionary
ISCO Code	ISCO Code (4 digits if possible)

2.8 Conclusion

Job portals might be a rich source of detailed information concerning two of the most critical variables for human resources analysis, which are the skills and the occupations required by employers. Nevertheless, to obtain consistent information for skills and occupational requirements from job advertisements, the use of dictionaries or classifications is needed, along with the implementation of more complex algorithms. Consequently, the first part of this section discussed and selected the best procedures to organise and categorise skills and occupational information.

First, for the Colombian case, information regarding skills is widespread in job advertisements. There is no national skills dictionary available to identify what words refer to in the job description for a certain skill; nevertheless, this section showed that the usage of international dictionaries such as the ESCO might facilitate building a methodology which identifies the skills demanded in each job advertisement for countries such as Colombia. Moreover, with the help of text mining techniques is possible to determine country-specific skills that are not listed in the ESCO dictionary, but are mentioned in the job vacancy description.

Second, job titles in vacancy advertisements can be, potentially, organised and coded into occupations. The categorisation of job titles into occupations is one of the most critical procedures because this variable summarises the main characteristics of labour demand (tasks and skills required), and this variable is a key input for other processes such as the imputation of wage and educational requirements. In this regard, the economic and statistic literature has developed different methods and algorithms to classify job titles into occupations (manual coding, classifiers, machine learning algorithms, etc.). Each method has advantages and disadvantages. Manual coding might ensure a relatively high level of accuracy (percentage of job titles coded correctly); however, given the large number of cases (job titles), manual classification is a time-consuming task. On the other hand, automatic coding might help to assign occupational codes over a relatively short period of time, but there might be a considerable number of observations misclassified. This accuracy rate depends on algorithm performance and database quality.

Among the automatic methods discussed in this section, there are two main statistical tools: machine learning algorithms and software classifiers (which contain a set of logic rules). The

main disadvantage of machine learning algorithms is that they strongly depend on the training database (job titles previously coded). In Colombia this kind of training database does not exist. Thus, software classifiers such as Cascot might be an excellent help in a context such as the Colombian one. However, Cascot does not successfully classify all the job titles.

Therefore, at least for the Colombian context, there is not a unique method that satisfactory assigns occupational code to the job titles. Given the advantages and disadvantages of each approach, this thesis proposes a combination of techniques: 1) manual coding for the most common job titles; 2) a software classifier (Cascot) adapted to the Colombian context, and, 3) an extension of a machine learning algorithm (nearest neighbourhood algorithm) that takes into account not only job titles but also skill requirements. Additionally, a (short) manual revision of the automatic outputs is undertaken.

Once all relevant variables are cleaned and adequately categorised for job vacancy analysis, another critical issue is the duplication problem. As vacancy data is collected from different websites (some of job advertisements can appear on more than one job board or even on the same job board) the second part of this section showed how to deal with duplicated records. Specifically, it was argued that a n-gram based approach (which is not sensitive to minor changes in string variables), so far, is the best method to minimise this issue. However, it is essential to recognise that (with the techniques available today) there is not a way (apart from using a time-consuming manual process) to demonstrate that all duplicated observations have been dropped.

Finally, relevant variables for the analysis of the labour demand for skills, such as wages and educational requirements, contain missing values. These missing values can create biases in the study of labour demand. Thus, the third part of this section explained and used the hot-deck and LASSO methods to impute missing values into the "education required" and "wage" variables.

In summary, this section 1) provided a robust and detailed methodology to obtain, organise and categorise skills and occupations from job portals for statistical analysis; 2) showed how to deal with duplicated job advertisements, and with missing values for relevant variables. Thus, as an outcome of this section and section 1, the vacancy database can now be tested.

3. References

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

Alexa (2017). *Website Traffic.* [online] Available at: https://www.alexa.com/siteinfo [Accessed 15 Oct. 2017].

Belloni, M., Brugiavini, A., Meschi, E., & Tijdens, K. (2014). *Measurement error in occupational coding: an analysis on SHARE data*. University Ca'Foscari of Venice, Dept. of Economics Working Paper Series No, 24.

Bethmann, A., M. Schierholz, K. Wenzig, and M. Zielonka. (2014). *Automatic Coding of Occupations.* In Proceedings of Statistics Canada Symposium. August 29–31, 2014, Quebec, Canada. Available at: http://www.statcan.gc.ca/sites/default/files/media/14291-eng.pdf (accessed October 01, 2018).

Burning Glass (2017). *The Digital Edge: Middle-Skill Workers and Careers.* September 2017, [online] Available at: https://www.burning-glass.com/wp-content/uploads/Digital_Edge_report_2017_final.pdf [Accessed 10 August. 2018].

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). Descriptive analysis of the vacancy database. Universidad del Rosario. Working Paper No. WP2-2020-004.

Cárdenas R., Jeisson. (2020c) Internal and external validity of the vacancy database. Universidad del Rosario. Working Paper No. WP2-2020-005.

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.

DANE. (2015). *Clasificación Internacional Uniforme de Ocupaciones. Adaptada para Colombia.* Available at: https://www.dane.gov.co/files/sen/nomenclatura/ciuo/CIUO_08_AC_2015_07_21.pdf. [Accessed 15 Jun. 2018]

ESCO. (2017). *ESCO handbook.* September 2017.

Green, F. (2011). *What is skill? An inter-disciplinary synthesis*. Centre for Learning and Life Changes in Knowledge Economies and Societies.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). *Three methods for occupation coding based on statistical learning.* Journal of Official Statistics, 33(1), 101-122.

ILO. (2008). *ISCO-08: Introductory and methodological notes.* International Labour Office.

ILO (2017). *ISCO*. [online] Available at: https://www.ilo.org/public/english/bureau/stat/isco/ [Accessed 23 Feb. 2017].

Jones, R., & Elias, P. (2004). *CASCOT: Computer-assisted Structured Coding Tool*. Coventry: Warwick Institute for Employment Research, University of Warwick, [Report]

Kankaraš, M., Montt, G., Paccagnella, M., Quintini, G., & Thorn, W. (2016*). Skills Matter: Further Results from the Survey of Adult Skills.* OECD Skills Studies. OECD Publishing.

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.

Lima, A., & Bakhshi, H. (2018). *Classifying occupations using web-based job advertisements: an application to STEM and creative occupations (No. ESCoE DP-2018-07)*. Economic Statistics Centre of Excellence (ESCoE).

Murphy, K. (2012) *Machine Learning A Probabilistic Perspective.* The MIT Press Cambridge, Massachusetts London, England.

Nigel, S (2016). *Webscraping for Job Vacancy Statistics.* Eurostat.

OECD (2017). *Latin American Economic Outlook 2017: Youth, Skills and Entrepreneurship.* OECD Publishing, Paris, https://doi.org/10.1787/leo-2017-en.

Oxforddictionaries (2017). *Definition.* [online] Available at: https://en.oxforddictionaries.com/definition/scrape [Accessed 21 May. 2017].

Reimsbach-Kounatze, C. (2015), *The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis.* OECD Digital Economy Papers, No. 245, OECD Publishing. http://dx.doi.org/10.1787/5js7t9wqzvg8-en

Spitz-Oener, A. (2006). *Technical change, job tasks, and rising educational demands: Looking outside the wage structure.* Journal of labor economics, 24(2), 235-270.

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

Varian, H. R. (2014). *Big data: New tricks for econometrics.* Journal of Economic Perspectives, 28(2), 3-28.

Wageindicator (2009). *EurOccupations: CASCOT software for coding job titles*. Available at: https://wageindicator.org/copy_of_documents/policy-briefs/European-Policy-Brief-no-3-CASCOT-coding-program-EUROCCUPATIONS-20100104.pdf [Accessed 10 Sep. 2018].

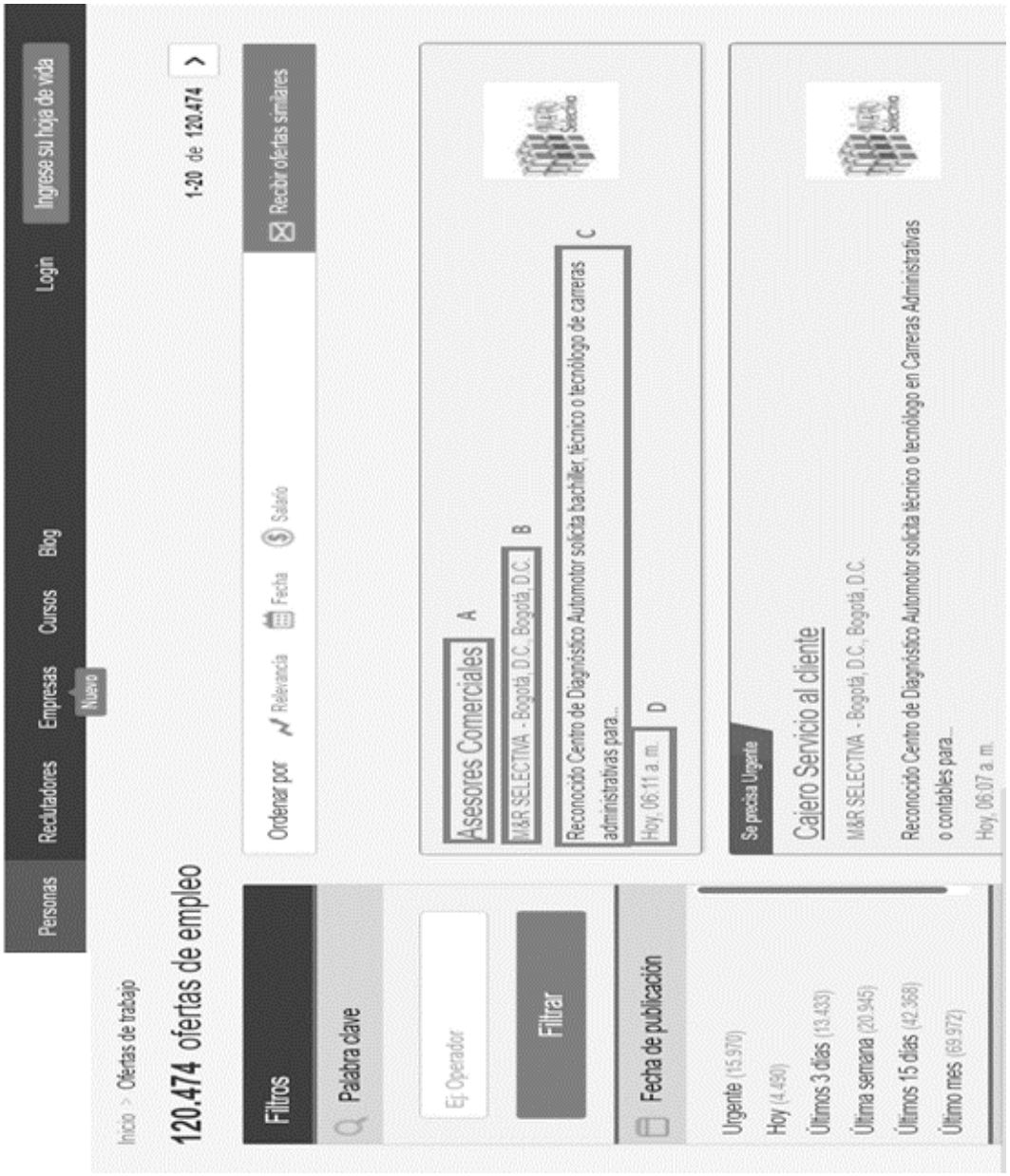
Appendix A: Examples of job portal structures

Figure A.1 shows how two websites ("Jobportal_a" and "Buscadordeempleo") present their vacancies differently. The red boxes in panel A and B highlight the job vacancy attributes that each website displays in listed job advertisements.

There are things in common between these two job portals: the red A boxes in the Jobportal_a and Buscadordeempleo panels highlight the job titles which are a short description about the position to be filled. However, there are also some differences between each website. Jobportal_a displays the name of the company that advertises the job, and the city where the vacancy (or vacancies) is available in Box B. In Box C a brief description of the job vacancy (e.g. level of education required by the employer) is shown, and Box D displays when the job was advertised. In contrast, in panel B, Buscadordeempleo displays information about the job title (Box A), the city where the vacancy is available (Box B), and the date when the vacancy is going to expire (Box C).

Figure A.1: Job portals comparison[41]

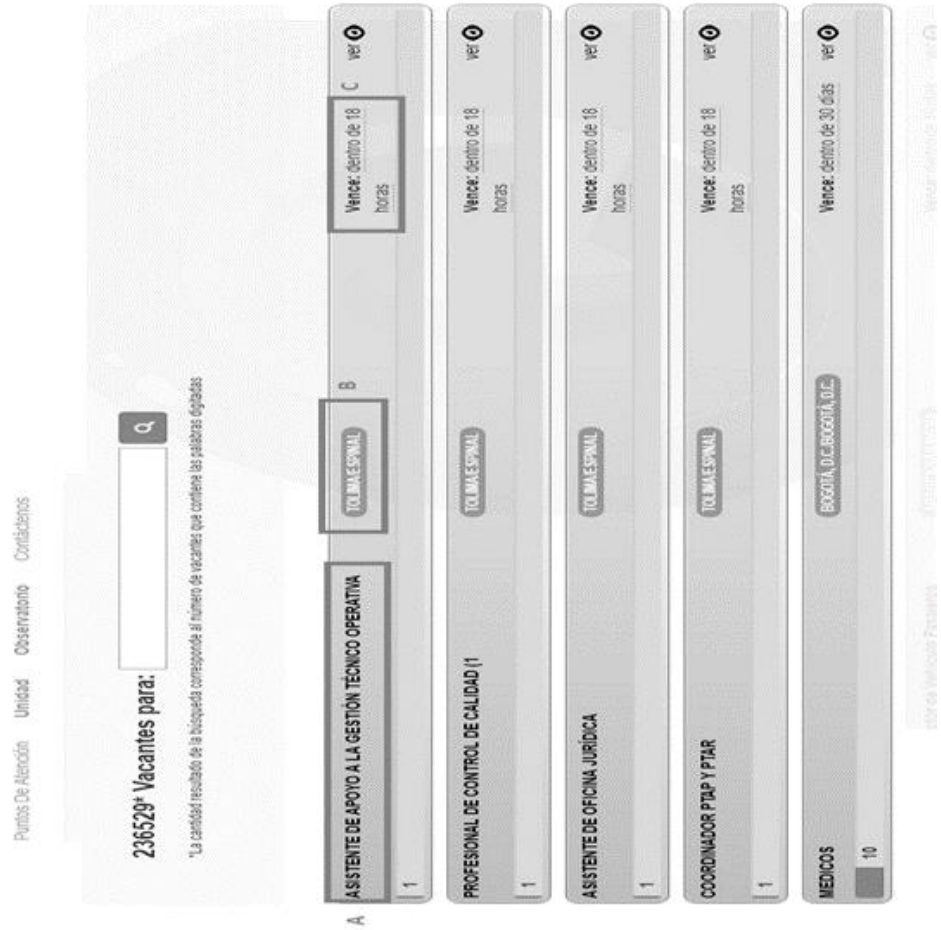
Panel A: Jobportal_a[42]



[41] The figures are presented in Spanish, because this is the original language used on the Colombian job boards.

[42] Translation of relevant parts of the text: Box A: Shopping assistant; Box B: Company's name: M&R selective, City: Bogotá; Box C: a recognized automotive diagnosis centre requires bachelors, technicians; Box D: Today, at 06:11 am.

Panel B: Buscadordeempleo[43]



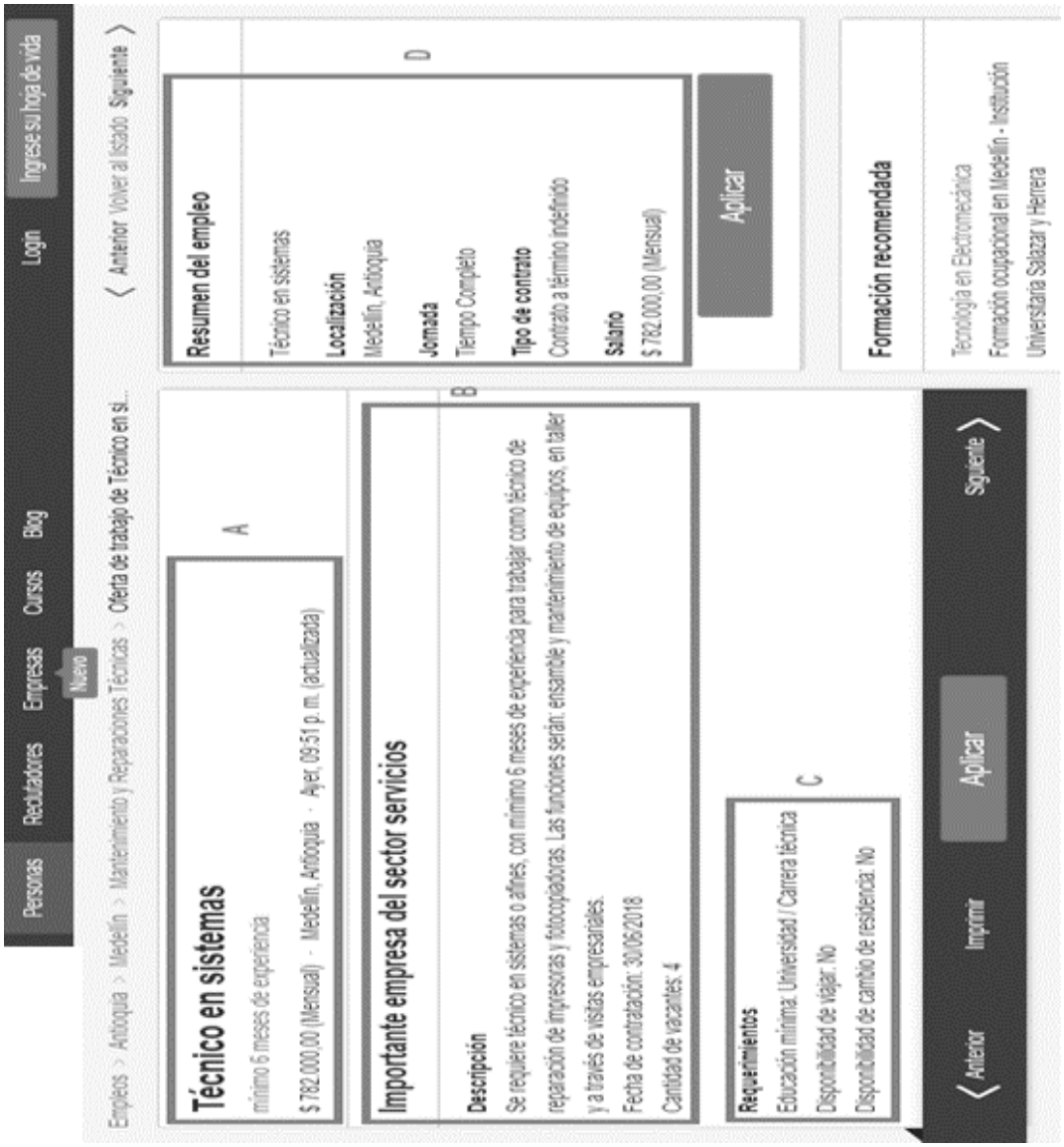
Sources: https://www.jobportal_a.com.co and buscadordeempleo.gov.co/

[43] Translation of relevant parts of the text: Box A: Assistant management operator; Box B: Department: Tolima, City: Espinal; Box C: Expiration date: in 18 hours.

Moreover, when more detailed information about each job is consulted, greater differences in how the information is presented arise between and within job portals. As observed in Figure A.2, information on the same website (in this case Jobportal_a) might vary from one advertisement to another. Panel A displays a job vacancy for a computer, automated teller, and office machine repairer, while Panel B requires a labourer and freight, stock, and material mover. Panel A shows information about the job title, experience required, wage offered, city and department (where the vacancy is available), and the date when the advertisement was published. In comparison, Box A in Panel B displays the job title, city and department (where the vacancy is available), and date when the advertisement was published. Consequently, information about required experience is not shown in Panel B. However, information regarding experience for this vacancy can be found in Box C (at the bottom of the website).

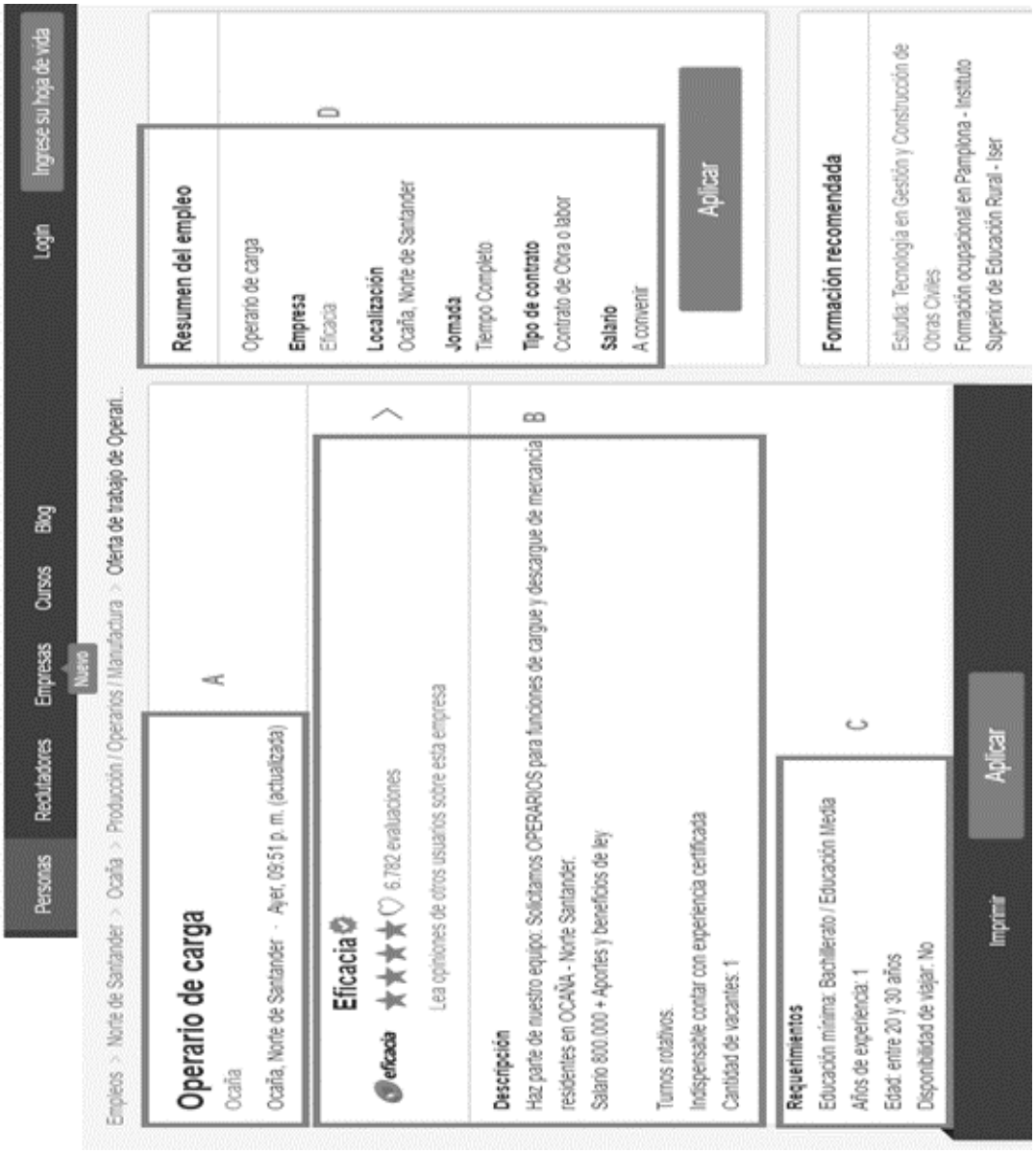
Figure A.2: Job advertisement comparison within the same job portal

Panel A: job one[44]



[44] Box A stands for: system support technicians. Minimum six months of work experience. Wage 782,000 pesos (monthly). City: Medellin. Department: Antioquia. Posted: yesterday at 09:51pm; Box B: Important company in the service sector. Description: System support technicians or similar are required with a minimum six months of work experience to repair printers and photocopiers. The tasks are: assembly and maintenance of equipment through business visits. Date of hire: 30/06/2018. Number of jobs: 4. Box C: Requirements. Minimum bachelor certificate required. No travel is required. Box D: Job summary. System support technicians. Localisation: Medellin, Antioquia. Working day: Full-time. Type of contract: indefinite term contract. Wage: 782,000 pesos (monthly).

Panel B: job two[45]



Source: https://www.jobportal_a.com.co

[45] Box A stands for: cargo operator. City: Ocaña. Department: Norte de Santander. Posted: yesterday at 09:51pm; Box B: Company's name: Eficacia. Description: Cargo operators are required to load and unload merchandise. Wage: 800,000 pesos (monthly). Rotating work shifts. Certificated experience is required. Number of jobs: 1. Box C: Requirements. Minimum bachelor certificate required. Years of experience: 1. Age: between 20 and 30 years old. No travel is required. Box D: Job summary. Cargo operator. Company's name: Eficacia. Localisation: Ocaña, Norte de Santander. Working day: Full-time. Type of contract: task type contract. Wage: to be agreed.

Panel A in Figure A.3 shows the HTML code of a Jobportal_a job advertisement. Information about the job title and company are delimited by the tags </p> and <h3>, respectively. In Panel B the HTML code for a job advertisement is displayed on “jobportal_c.gov.co”. On this website, the job title is defined by the syntaxis “<h2>” and the company information by the syntax <h4>Empresa:.

Figure A.3: Code comparison between job portals

Panel A: Jobportal_a

Panel B: Buscadordeempleo

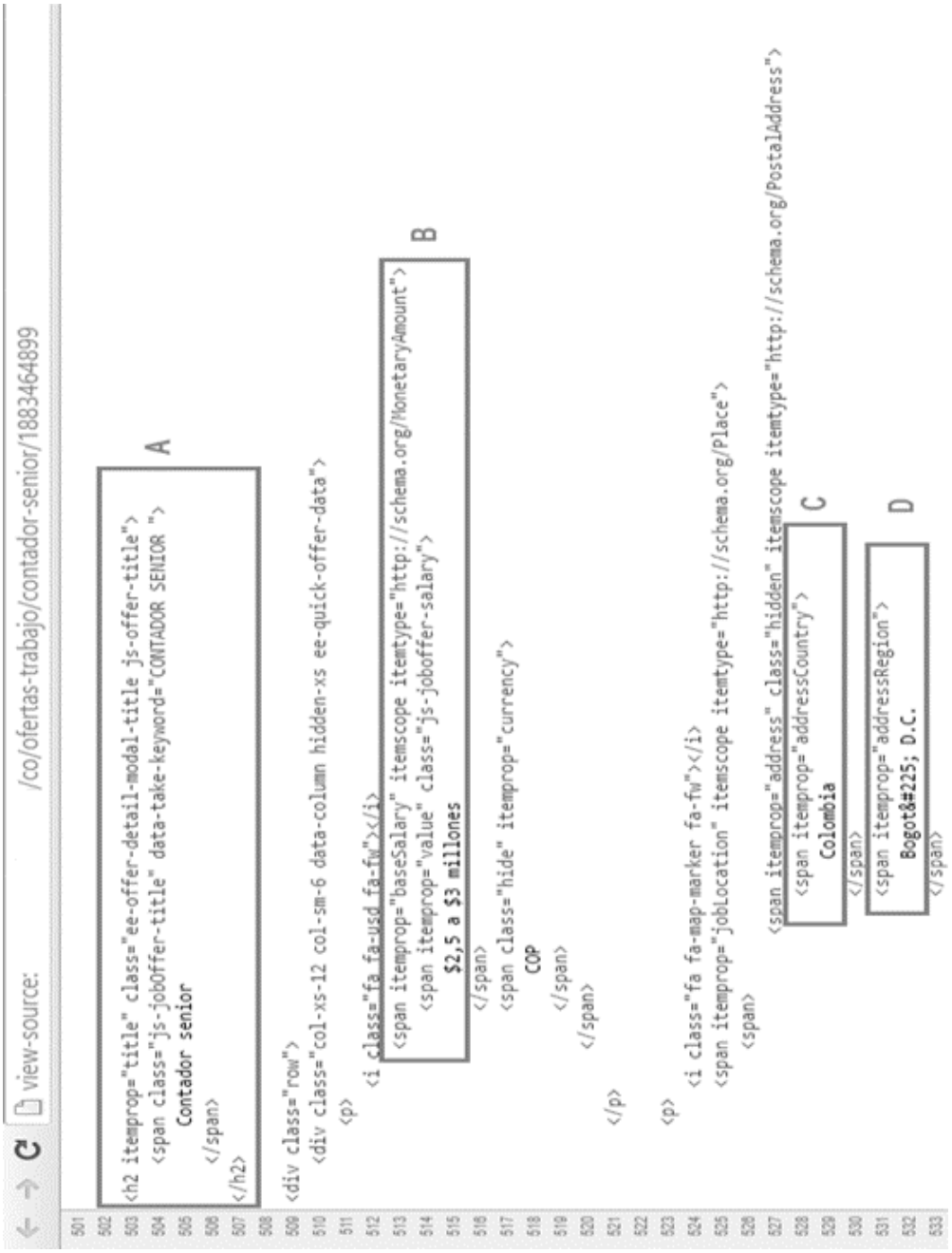
Sources: https://www.jobportal_a.com.co and buscadordeempleo.gov.co/

For illustrative purposes, Figure A.4 shows the HTML code structure of the web page Jobportal_b for a job advertisement. The web scraping technique recognises the tags (or Xpath) where relevant information exists. In Box A, the tag '"class='js-jobOffer-title'" contains the information related to a job title ("*Contador Senior*"); in Box B, the tag "class=='js-jobOffer-salary'" ($2,5 a $3 *millones*) describes the salary offered; and the tags "itemprop='addressCountry'" and itemprop='addressRegion'" in Box C and D, respectively, contain information regarding the country and the region where the vacancy is offered (Colombia and Bogotá, respectively[46]). Consequently, the R codes (developed in this thesis) recognise each one of those tags (or Xpath) and extract the information of interest for each job advertisement from each job portal.

It is important to note that sometimes companies do not advertise information regarding a characteristic of a vacancy such as salary using the corresponding tag (e.g. "class=='js-jobOffer-salary'"). In these cases, the algorithm searches for the tag, and when the tag is not found, the algorithm leaves a missing value in the database. This issue does not necessarily mean that there is no information regarding a certain job characteristic (e.g. salary). Employers might have posted the job characteristics using other tags, for instance, in the job description tag ("class=='offer-detail'"). Indeed, it is common to see that employers post most of the relevant information using the tag description of the vacancy (which is a paragraph where companies describe the job position), while other tags, such as for salary, might not be used. Consequently, as will be seen in Appendix B:, the implementation of text mining techniques is necessary to identify all the relevant information in job advertisements.

[46] As shown in Chapter 7, as expected, most of the vacancies are available in Colombia. However, there are some offers to work in other countries.

Figure A.4: HTML code structure



Source: https://www.jobportal_b.com

Appendix B: Text mining

Column 1 and 2 of Table B.1 shows part of the information provided by employers for a pair of job vacancies. In the job description (Column 1) for the first vacancy, the employers mention that a person is required with an undergraduate certificate; however in the website column, where the employer was supposed to provide specific information regarding educational requirements (Column 2), there is a missing value. In contrast, for the second vacancy (see the second row of Table B.1), in the job description (Column 1) the employer does not mention any educational requirements. Indeed, for this vacancy the information regarding education is available in the "Educational requirements" column (Column 2).

Thus, the algorithm needs to "read" the different columns of the scraped data (not only the educational requirements column) to identify qualification requirements. Some of the relevant information might be only mentioned in the job description or in other specific columns; additionally, information might be repeated in different columns. In this example, the algorithm creates "Dummy_Undergraduate_certificate" and "Dummy_PhD_certificate" columns (the third and fourth columns of Table B.1). Based on the information provided in the job description, the "Dummy_Undergraduate_certificate" column identifies (takes a value of 1) that the first vacancy required a person with an undergraduate certificate while the second vacancy does not require a person with an undergraduate certificate. On the other hand, based on the "Educational requirements" column, the "Dummy_PhD_certificate" column identifies that the second vacancy (second row of Table B.1) requests a person with a PhD. It is important to note that employers might be indifferent about educational levels or another job characteristic. For instance, a vacancy might require a person with a high school or undergraduate level. In these cases, the dummy variables ("high school" and "undergraduate_certificate") take the value of 1 at the same time.

Table B.1: Example of the content of a scraped database

Job description	Educational requirements	Dummy_ Undergraduate certificate	Dummy_ PhD certificate
• Must have some previous production experience in a book publishing or printing environment • Excellent communication and interpersonal skills and consultative customer care approach • **Undergraduate certificate** • Strong IT skills with experience of using a CRM system advantageous	No information provided by the employer (missing value)	1	0
• Candidates with expertise in cancer genomics and related disciplines, and candidates with experience of working with clinical data, are particularly encouraged to apply.	A PhD (or equivalent) in a field related to cancer and/or genomics and significant research experience.	0	1

• Editorial experience is not required, although applicants with significant editorial experience are encouraged to apply and will potentially be considered for a Senior Editor position.			

Additionally, there is an issue when looking for patterns in the Spanish language. In this language, nouns have a gender; for instance, an undergraduate might be called "*universitario*" (for men) or "*universitaria*" (for women). Given this fact, and the usage of synonyms to express a job requirement, the algorithm looks for patterns in the root of the words[47]. The selection of the root of the words is a critical process where it is necessary to carefully select the proper roots to correctly classify job requirements. In this way, the dummy variables created are guaranteed to correctly identify employers' requirements, even with the presence of synonyms, nouns with genders, etc.

[47] For instance, in the undergraduate case, the algorithm looks for "universi" among other word roots.

Appendix C: Detailed process description for the classification of companies

C.1. Manual coding

Manual coding is a process through which a person (or a group of people) manually assign a code (or category) to each observation based on the data's characteristics. Similar to the coding of job titles (see Chapter 6), the full manual coding of each company's sector is a time-consuming task. There are a few companies (most of them related to office administration, office support, and other business support activities) that post relatively more vacancies than others. Consequently, a manual coding process was conducted to ensure that the most frequently used versions of companies' names were properly classified. Fifty companies' names were coded manually[48]. These companies represent around 13% of the total number of companies listed in the vacancy database. Once this process was completed, the next step was the implementation of automatic coding using word-based matching methods.

C.2. Word-based matching methods ("Fuzzy merge")

Different word-based matching methods exist. The differences between one method and another are due to the rules employed to obtain a similarity score. For instance, the Levenshtein distance algorithm (Levenshtein, 1996) calculates the distance (similarity) between two words or sentences (strings) based on the minimum number of characters that are necessary to change one word (or phrase) into the other word (or sentence). As an illustration, when the Levenshtein distance algorithm compares two words such as "butcher" and "butchery" the distance will be 1, which is the number of characters required to transform the word "butcher" into "butchery". Moreover, other algorithms are available, such as Soundex, in which metrics are based on the sound of the words rather than the characters of the words (see Holmes and McCabe 2002, for a detailed explanation of the Soundex algorithm).

Given the numerous algorithms available, this thesis used the following steps to merge companies' names for each job and with the information available in the RUES database. First, observations were merged that shared the same names in the vacancy and the RUES database.

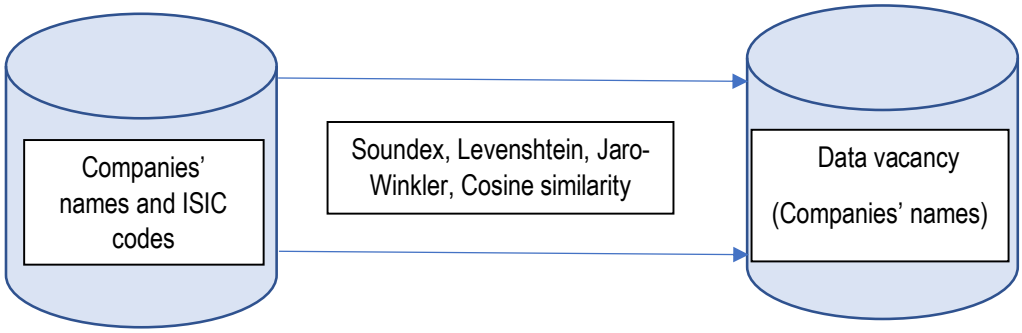
[48] In the case of multi-product/sector companies, the RUES only registers the main activity reported by the company. Consequently, the ISCO code assigned to those companies corresponds to the main economically activity reported in the RUES.

Only 2% of companies in the vacancy database were matchable with their ISIC code. As mentioned above, this low merge rate might be due to differences between companies' names in the two databases. Consequently, it was necessary to use fuzzy merge algorithms to correctly assign an ISIC code to the maximum number of companies' names efficiently. More specifically, the Jaro-Winkler algorithm was used to merge the databases (see Cohen et al. 2005, for a detailed explanation of the Jaro-Winkler algorithm). A threshold of 98%[49] of similarity was selected. With this method, 15% of the vacancy database received an ISIC code. For the remaining database, a Cosine similarity algorithm was implemented at a 98% matching threshold (see Huang 2008, and Appendix G: for a detailed explanation of the Cosine similarity algorithm). By doing so, a further 9% of observations in the vacancy database were assigned an ISIC code. Additionally, the Soundex algorithm at a 98% matching threshold was executed. With this procedure, 6% of observations in the vacancy database were assigned an ISIC code.

Finally, the Levenshtein algorithm was implemented. Given the characteristic of this algorithm, that it is sensitive to the length of words, it was executed only for those observations that had not been merged and for which the length of a company's name (number of characters) was more than 5. With this algorithm, 3% of job announcements were merged with the RUES database. Therefore, 48% of observations in the vacancy database (manually and with word-based matching methods) received an ISIC code, up to this point. As can be observed, despite these word-based matching methods, there is still a considerable number of observations without an ISIC code (52%). This outcome might be due to important differences between companies' names in the vacancy and the RUES database. Or the issue might reside in the RUES database, which a significant number of companies might not be registered with.

[49] It is important to note that matching thresholds were relatively high in order to guarantee an acceptable accuracy level.

Figure C.1: Fuzzy merge: the classification of companies



C.3. A return to manual coding

As a considerable number of observations (52%) were not coded with word-based matching methods. In order to code most of the vacancies, it was necessary to return to a manual coding process. As in Subsection C.1, first, coding was carried out via a visual inspection of uncoded information on the vacancy database, e.g., the companies that were not coded by using the word-based matching methods and the manual coding methods mentioned in the previous subsection. Subsequently, those companies' names that appeared more frequently in the database were manually coded: a total of 50. Yet, these companies represent 4% of the vacancy database; therefore, even with the above procedure there a considerable number of observations without an ISIC code (48%). Thus, and as the last step, the companies' names were used to assign ISIC codes. Frequently, companies' names reveal the sector where they perform their activities. This is the case, for instance, of restaurants (McDonald's restaurant) and universities (University of Santander), among others. Consequently, by using keywords from company names is possible to assign an ISCO code to a considerable number of job announcements. With this last procedure, it was possible to assign an ISIC code to 9% of the vacancy announcements.

Appendix D: Machine learning algorithms

To assign occupational codes, the basic machine learning model starts by transforming job titles into numeric values. A variable x_{ij} takes the value of 1 if the word j occurs in document i. By applying the above transformation to all X documents in a database, the result is a word co-occurrence matrix that indicates which words appear in each document. For instance, as shown in Table D.1, one job title might be "Web designer", while another would be "Network designer". Thus, three indicating variables ("web", "network" and "graphic") will be created that take values of 1 if the word appears in the job title and 0 otherwise.

Table D.1: N-grams based on job titles

ISIC code	Job title	Web	designer	Network
2166	Web designer	1	1	0
2523	Network designer	0	1	1

This "bag of words" representation is a key element for the algorithm to learn how to classify job titles into occupations. Moreover, the computer requires an algorithm to classify job titles into occupational codes based on this bag of words. There are different kinds of algorithms—such as linear and logistic regression, random forest, naïve Bayes, among others—that help to map x inputs into the outputs y (where $y \in \{1 \dots, C\}$). The choice between one algorithm or another depends on the problem to be solved and the available data.

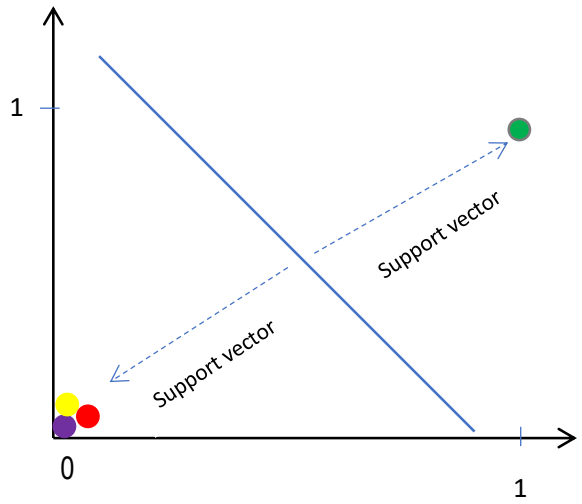
Appendix E: Support vector machine (SVM)

SVM is a machine learning algorithm. Briefly, the SVM algorithm based on the inputs (x) tries to find a hyperplane(s) (which, for simplicity, can be thought of as a line in a Cartesian plane) that best divides the data into two or more classes. In the case of job titles, the SVM algorithm considers the numeric transformation of the job titles (x_{ij} variables) to create the hyperplane(s) that separates the data in different occupational groups. For instance, Figure E.1 shows a simple illustration of how SVM works to classify a job title such as "gardener" into an occupational code. In this example, a vacancy database has an observation listed with the job title "gardener"

(occupational code 6113—Gardeners; Horticultural and Nursery Growers), while the job titles in other observations are “account” (4311—Accounting and bookkeeping clerks), “economist” (2631—Economists) and “psychologists” (2634—Psychologists) (indicated by the purple, yellow and red points, respectively, in Figure E.1). For simplicity, x input (x-axis) is the numeric transformation of the word “gardener” (which takes a value of 1 if the word “gardener” occurs in vacancy advertisement i).

Moreover, the outcome y takes value of 1 if the occupational group is “6113—Gardeners; Horticultural and Nursery Growers” (green point in Figure E.1). In this example, it is relatively easy to find different hyperplanes that have divided the dataset into two parts (numbers of classes). However, the best hyperplane that separates the data is the one that maximises margins between the points (0,0) and (1,1) on the Cartesian plane. These two points are named the support vectors, which are the points nearest to the hyperplane. Consequently, the greater the distance between the margins and the hyperplane, the higher the probability of correct classification. Thus, in this way, the computer learns how to classify job titles into occupations. Clearly, the algorithm needs to do more complex tasks when the data contains a significant number of observations, classes and explanatory variables (x).

Figure E.1: SVM classification with job titles



Appendix F: SVM using job titles

As mentioned above, the basic machine learning approach uses job title information to assign occupational codes. This section evaluates the possibility of predicting the remaining job titles of the Colombian database by conducting an underlying machine learning approach with the SVM method. The job titles categorised in the previous subsections (6.4.1 to 6.4.6) were used to train and test the algorithm. Additionally, the cleaned job titles (Subsection 6.4.2) were transformed into a word co-occurrence matrix with around 4,000 terms. As mentioned above, the algorithm uses this bag of words as an input (x) to classify job titles into occupations. Seen via a manual check, 40% of job titles were incorrectly classified. Therefore, the SVM machine learning algorithm which only uses job titles is not an option to classify the remaining observations in the vacancy database.

Appendix G: Nearest neighbour algorithm using job titles

There are different ways to measure the distance between two strings (see Appendix C:), however, as Gweon et al. (2017) note, one of the most common approaches is the Cosine similarity. This approach takes the vector representation of two documents (e.g. a and b) and calculates the distance between vectors in the following way:

$$Similarity(A,B) = \frac{\mathrm{A}*\mathrm{B}}{||\mathrm{A}||*||\mathrm{B}||} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

In the above formula, A and B are the vector representations of document a and b. This similarity score can lie between 1 and 0. The greater the similarity between the two documents the closer the value is to 1. For instance, Tables G.1 and G.2 are examples of a training dataset with a new observation to be coded. In both tables the new record to be coded is “Web designer”.

The training database in Table G.1 has a unique observation in which the job title is “Web designer” and this title has been assigned with ISCO code “2166” (Graphic and multimedia designers). Consequently, the vector representation of the training database and the new record is shown in columns 4 and 5 of Table G.1.

Table G.1: Vector representation example one

Source	ISCO code	Job title	Web	designer
Training dataset	2166	Web designer	1	1
New record	…	Web designer	1	1

Consequently, the similarity score between the two records for "Web designer" in Table G.1 is given by the following formula:

$$Similarity(new\ vacancy, row1) = \frac{(1 * 1) + (1 * 1)}{\sqrt{1^2 + 1^2} * \sqrt{1^2 + 1^2}} = 1$$

In contrast, the training database in Table G.2 is composed of a unique observation in which the job title is "Network designer" and the ISCO code "2523". Consequently, the vector representation of the training database and the new record is shown in columns 4 and 5 of Table G.2.

Table G.2: Vector representation example two

Source	ISCO code	Job title	Web	designer	Network
Training dataset	2523	Network designer	0	1	1
New record	.	Web designer	1	1	0

The similarity score between the two records is given by:

$$Similarity(new\ vacancy, row2) = \frac{(0*1)+(1*1)+(1*0)}{\sqrt{1^2+1^2}*\sqrt{1^2+1^2}} = 0.5 \qquad (2)$$

Therefore (as expected), the most similar observation in the training database for a new entry with a job title of "Web designer" is the one with the same string values. The new record would receive the occupational code 2166 (Graphic and multimedia designers). As Gweon et al. (2017) argue, the reason for the higher accuracy of the nearest neighbour algorithms compared to the

SVM algorithms is that the former are more effective when the accuracy of the result highly depends on local points (very similar job titles), and little information can be obtained from remote records (dissimilar job titles).

In this regard, Gweon et al. (2017) propose an improvement using the Cosine similarity score in the following way: they denote a new document (job title) as x, the number of the nearest neighbours in the training dataset as K(x) and s(x) the (cosine) similarity score of the nearest neighbours (job titles). Additionally, $k_i(x)$ out of the K(x) neighbours have the class (occupational code) $c_i(i = 1,2,3, \ldots \mathrm{L})$. Consequently, the rule to assign an occupational code is defined as:

$$\gamma(c_i|x) = p(c_i|x)s(x)\left(\frac{K(X)}{K(X)+0.1}\right) \quad (3)$$

As Gweon et al. (2017) note, the predicted code only depends on $p(c_i|x)$. The terms $K(X)$ and $s(x)$ are constant for any observation in the training database with one string in common with the new record. Both $K(X)$ and $s(x)$ terms help to identify which new records have enough and similar neighbours to make a proper comparison. The multiplier $s(x)$ indicates the degree of closeness, a high value of $s(x)$ indicates that the job titles in the training database are very similar to the new record(s). Moreover, the term $\left(\frac{K(X)}{K(X)+0.1}\right)$ is a control for the number of neighbours. The higher this indicator, the larger the number of nearest neighbours for the new record(s). Consequently, it is supposed that the algorithm's output will be more precise when there are more nearest neighbours ($K(X)$). As Gweon et al. (2017) mention, at most this multiplier can reduce the γ score by about 10% (when $K(X)$=1), while $p(c_i|x)$ and $s(x)$ can lower γ to zero.

Consequently, for this algorithm, the term $\left(\frac{K(X)}{K(X)+0.1}\right)$ has relatively less importance than the other terms. The constant 0.1 serves to decrease the importance of this term in total score γ. The choice of 0.1 might be seen as arbitrary. However, a smaller constant makes the γ score more sensitive to changes in $K(X)$ which are not desirable due to the other two terms ($s(x)$ and $p(c_i|x)$) which are relatively more important for this algorithm. On the other hand, a larger constant makes the γ score less sensitive to changes in $K(X)$ which would make irrelevant the term $\left(\frac{K(X)}{K(X)+0.1}\right)$.

Finally, the class in which the γ score has the highest values will be assigned to the new record. Table G.3 illustrates how this algorithm works for a given training database. There is a new record with the following words in the job title "technician assistant food". There are four other observations in the training database that have one word in common with the new record. Columns 2, 3 and 4 (of Table G.3) show the vector representation of the training database and the new record. Three-quarters of the observations in the training database coincide with the new record due to the word "food". These observations have the occupational code "9412" ("Kitchen helpers"). Additionally, another observation in the training data set coincides with the new record due to the word "assistant" and has the occupational code "5223" referring to "Shop sales assistants". Following equation 3, the values for $p(c_{9412}|x)$ and for $p(c_{5223}|x)$ are 0.75 (3/4) and 0.25, respectively. Thus, the observations with the occupational code 9412 receive the same $p(c_i|x)$ value (0.75), while the observation with the occupational code 5223 receives the value 0.25 (Column 6 of Table G.3).

For each observation in the training database, the $s(x)$ (Cosine similarity) with the new record is given by $\frac{1}{\sqrt{1}\sqrt{3}} = 0.5774$ (Column 7). Finally, the term $\left(\frac{K(X)}{K(X)+0.1}\right) = \left(\frac{4}{4+0.1}\right)$ is equal to 0.9756 for each observation (Column 8 of Table G.3). By multiplying the terms, $\gamma(c_{9412}|x)$ is equal to 0.4225, while the $\gamma(c_{5223}|x)$ score is equal to 0.1408 (Column 9). Thus, the occupational code assigned to the new record "technician assistant food" is 9412 ("Kitchen helpers") (Column 5 of Table G.3).

Table G.3: Nearest neighbour algorithm (Gweon et al. 2017)

Source	technician	assistant	food	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	9412	0.75	0.5774	0.9756	0.4225
	0	0	1					
	0	0	1					
	0	1	0	5223	0.25			0.1408
New record	1	1	1	9412*				

* Final occupational code assigned to the new record

However, when the training database is unbalanced, namely, there some job titles that are significantly more common than others (for instance shop assistant for the Colombian case) this algorithm might fail. Table G.4 shows an example of this issue. Supposing that there is training a database with four observations, the job title of one of those observations is "help preparation food", and the job title of the remaining observations in the training database is "shop sales assistant". Additionally, there is a new record with the following words in the job title "technician assistant food". Columns 2 to 8 of Table G.4 show the vector representation of the training database and the new record. Three-quarters of the observations in the training database coincide with the new record due to the word "assistant". These observations have the occupational code 5223 ("Shop sales assistants"). Additionally, another observation in the training data set coincides with the new record due to the word "food" and has the occupational code "Kitchen helpers" (9412 ISCO) (Column 9 of Table G.4). Following the equation 3, the values for $\mathrm{p}(\mathrm{c}_{9412}|\mathrm{x})$ and for $\mathrm{p}(\mathrm{c}_{5223}|\mathrm{x})$ are 0.25 (1/4) and 0.75, respectively (Column 10).

For each observation in the training database, the $\mathrm{s}(\mathrm{x})$ (Cosine similarity) with the new record is given by $\frac{1}{\sqrt{3}\sqrt{3}} = 0.33$ (column 11 of Table G.4). Finally, the term $\left(\frac{\mathrm{K(X)}}{\mathrm{K(X)}+0.1}\right) = \left(\frac{4}{4+0.1}\right)$ is equal to 0.9756 for each observation (Column 12 of Table G.4). By multiplying the terms, $\gamma(\mathrm{c}_{9412}|\mathrm{x})$ is equal to 0.0812, while the $\gamma(\mathrm{c}_{5223}|\mathrm{x})$ score is equal to 0.2438 (Column 13). Thus, the occupational code assigned to the new record "Shop sales assistant" is 5223 (Column 9 of Table G.4). Consequently, when the training database is unbalanced (there are many job titles with the same word), the algorithm might classify all the new records that contain the word "assistant" in the category of "shop sales assistants" (ISCO code 5223). Consequently, the accuracy level of the algorithm decreases. There are observations such as "technician assistant food" that do not receive the proper occupational code.

Table G.4: Limitation of the nearest neighbour algorithm

Source	technician	assistant	food	Preparation	help	sales	shop	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	1	1	0	0	9412	0.25	0.333	0.9756	0.0812
	0	1	0	0	0	1	1	5223	0.75			0.2438
	0	1	0	0	0	1	1					
	0	1	0	0	0	1	1					
New record	1	1	1	0	0	0		5223*				

* Final occupational code assigned to the new record

One possibility to avoid this issue is to perform a resample of the training database to balance the occupational groups. However, this approach might not be effective. Even when two observations with the word “assistant” are dropped the predicted result for the new record will be the same as before (“shop sales assistants” ISCO 5223). Additionally, this re-balance affects the occupational structure of the training database and some job titles that before were predicted correctly. With this adjustment, some of them might be misclassified.

Tables G.5 and G.6 illustrate an example of this method. For the new record “technician assistant food”, there are six observations in the training database that have one word in common in the job title. With this information, the algorithm proposed in Gweon et al. (2017) would incorrectly code the new record in the “shop sales assistants–5223” category, as shown in Table G.5.

Table G.5: An extension of the nearest neighbour algorithm (part 1)

Source	Job title		Job description (skills)					Parameters			
	technician	assistant	food	dispose waste	use food cutting tools	clean surfaces	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	0	1	1	9412	0.333	0.5774	0.983	0.189
	0	0	1	1	0	1	9412		0.5774	0.983	0.189
	0	1	0	1	0	0	5223	0.5	0.5774	0.983	0.283
	0	1	0	0	0	0	5223		0.5774	0.983	0.283
	0	1	0	0	0	0	5223		0.5774	0.983	0.283
	1	0	0	0	0	0	3112	0.166	0.5774	0.983	0.094
New record	1	1	1	1	1	1	5223*				

* Final occupational code assigned to the new record

However, considering the skills information being demanded helps to identify a more precise sample of nearest neighbours in this example. More specifically, from Table G.5 it is possible to note that in the new record employers demand some skills, such as “dispose waste”, “use food cutting tools” and “clean surfaces”. Moreover, in the training database those skills are also demanded. The skills “use food cutting tools” and “clean surfaces” were mentioned for the first row, while the skill “dispose waste” was mentioned in the third row. Consequently, it is possible to drop all those observations in the training database which do not have skills in common with the new record. By doing so, the last 3 rows in the training dataset are dropped (see Table G.6). As a result, Table G.6 presents a more precise training base for the algorithm. Indeed, the predicted code for the new record is “Kitchen assistant–9412” which seems to be more accurate than “Shop sales assistants–5223”.

Table G.6: An extension of the nearest neighbour algorithm (part 2)

Source	Job title		Job description (skills)					Parameters			
	technician	assistant	food	dispose waste	use food cutting tools	clean surfaces	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	0	1	1	9412	0.75	0.5774	0.967	0.418
	0	0	1	1	0	1	9412				
	0	1	0	1	0	0	5223	0.25	0.5774	0.967	0.139
New record	1	1	1	1	1	1	9412*				

* Final occupational code assigned to the new record

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 Cientfica-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.