5. Methodology

Jeisson Cárdenas Rubio
5. Methodology
5.1. Introduction

The analysis of labour demand information is a relevant factor to improve people’s skills according to employer requirements. As mentioned by the OECD (2017b), 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 book, unemployment, informality, etc. However, as discussed in Chapter 4, this capacity to analyse labour demand, in most countries, has been hampered by a lack of information about employer requirements.

Recently, online job portals have caught the attention of researchers and policymakers insofar as they might fill the labour demand information gap (Kureková, Beblavy, and Thum, 2014; Reimsbach-Kounatze 2015). These job portals contain a large number of job advertisements, which are accessible to anyone interested in vacancies and employer 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 no one 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, vacancy information is not organised, categorised, and consolidated in a database for statistical purposes. Thus, it is necessary to develop a robust methodology that collects, organises, categorises,
and analyses labour demand using job portals. This chapter proposes and explains each of these methodological steps. The second section of this chapter describes what information is available from Colombian job portals. The third section analyses the most important and reliable job portals to investigate how to conduct a proper labour demand analysis. Given that there is a huge amount of vacancies available online and, consequently, the manual collection of labour demand information is virtually impossible, the fourth section describes web scraping techniques that can be used to automatically collect online job advertisements. The fifth section explains the organisation (homogenisation) of different information from job portals into a single database 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,” it is necessary to implement other and more complex text mining techniques; thus, the last section of this chapter shows how it is possible to identify the sector where the employer belongs.

5.2. Measurement of the labour demand: Job vacancies

As mentioned in more detail in Chapter 2, a job vacancy can be understood as a vacant position within a company that is trying to fill it. Companies recruit potential workers in diverse ways to fill their vacancies. Likewise, as discussed in Chapter 4, 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 overrepresented; see Chapter 4), they might capture a large share of the net and replacement labour demand behaviour.
Table 5.1 shows the most important job portals in terms of data traffic (number of visitors) available in Colombia (Alexa 2017). For instance, “https://www.jobportal_a.com.co/” is the 37th most visited web page in Colombia, while “https://www.jobportal_b.com/” is the 89th. Additionally, Column 3 in Table 5.1 shows the number of job advertisements available for each job portal in October 2017.

Table 5.1. Average number of job advertisements and traffic ranking for selective Colombian job portals

<table>
<thead>
<tr>
<th>Colombia</th>
<th>Alexa Rank</th>
<th>Number of job advertisements</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.jobportal_a.com.co/">https://www.jobportal_a.com.co/</a></td>
<td>37</td>
<td>115,723</td>
</tr>
<tr>
<td>https://www.jobportal_b.com/</td>
<td>89</td>
<td>62,732</td>
</tr>
<tr>
<td><a href="https://www.jobportal_c.gov.co/">https://www.jobportal_c.gov.co/</a></td>
<td>199</td>
<td>263,621</td>
</tr>
<tr>
<td><a href="https://www.jobportal_d.com.co/">https://www.jobportal_d.com.co/</a></td>
<td>1,015</td>
<td>172,440</td>
</tr>
<tr>
<td><a href="https://www.jobportal_e.com.co/">https://www.jobportal_e.com.co/</a></td>
<td>2,280</td>
<td>20,143</td>
</tr>
<tr>
<td><a href="https://www.jobportal_f.com.co/">https://www.jobportal_f.com.co/</a></td>
<td>3,683</td>
<td>46,853</td>
</tr>
</tbody>
</table>


A job advertisement is understood as a text on a job portal that shows relevant information about a job vacancy (Swier 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 5.1 is that there is a large number of job advertisements and job vacancies on each website. 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, copying the information and pasting it in a database thousands of times.

---

48 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.

49 This book anonymised the job portals (removed their names and web page address) in order to protect their identity and to avoid promoting a particular job portal.

50 In Colombia, companies advertise their vacancies on different websites and, depending on the job portal, the cost of promoting a vacancy varies between £24 and £26.
Each job portal shows a list of available vacancies. Nevertheless, each website organises and shows its data according to its own criteria (see Appendix A, Figures A.1 and A.2). Table 5.2 (below) summarises the differences between two job advertisements within the same job portal. Even though this website presents almost the same information about the two vacancies, the localisation and categorisation of these variables (such as experience and wages, among others) might vary according to the 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 5.2. Consequently, a job portal is a semi-structured source of vacancy information. This feature makes it difficult to automatically collect data from these website sources. Thus, an algorithm that collects this information needs to recognise differences between advertisements and organise the information in order to properly construct/calculate totals for the database of net and replacement labour demand (hereinafter labour demand database; see Chapter 8).

Table 5.2. Job advertisement structure comparison within the same job portal

<table>
<thead>
<tr>
<th>Variables</th>
<th>Panel A: First job advertisement</th>
<th>Panel B: Second job advertisement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Box A</td>
<td>Box B</td>
</tr>
<tr>
<td>Job title</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Experience</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Wage</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Location</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Publication date</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Company name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Number of jobs</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Educational requirement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of contract</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Required age</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Jobportal_a.
Differences between job announcements also arise when comparing two different websites. For instance, Figure 5.1 compares two job advertisements: one posted on Jobportal_a (Panel A) and the other posted on Jobportal_c (Panel B). Both advertisements require an “accountant” (see Box A, Panel A and Panel B) (note that this is not the same vacancy posted on different websites); nevertheless, 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 Box C (at the bottom of Panel A) and Box D (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 years. In contrast, for Jobportal_c, the wage variable is expressed in ranges based on the current legal minimum wage\(^{51}\) and the experience variable is shown in terms of months.

Even though these formatting 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, 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), etc.—; additionally, 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).\(^{52}\)

\(^{51}\) In Colombia, every year, the national government decrees the minimum remuneration for a full-time job (current legal monthly minimum wage). For the year 2018, the minimum wage was 781,242 Colombian pesos (around £196) per month.

\(^{52}\) 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>.”
Figure 5.1. **Job advertisement comparison between job portals**

Panel A: Jobportal_a

---

This heterogeneity between and within websites makes it difficult to automatically collect information. 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 the website and each job announcement. Thus, in order to collect as much information as possible on labour demand, 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.
5.3. Selecting the most important vacancy websites in the country

As shown above, there exist different websites with relatively high data traffic (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. First, 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, and the effort and time to collect information from the selected websites increases significantly as a result.

Second, when considering a larger number of portals, duplication problems arise (as discussed in Chapter 6 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 Chapter 6), the probability of duplication persists and even increases by adding more websites. Yet, if a single job portal is used to build a vacancy database, other issues arise. 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 there exist relevant sources of job vacancy information and computational capabilities, to decrease the possible bias of utilising one source, it is necessary to consider job advertisements from different websites to build a vacancy database. In order to select the job portals that best capture the dynamics of the Colombian labour market, the following criteria were applied: 1) volume (number of advertisements available), 2) website quality (structure and number of variables or granularity of information), and 3) traffic ranking (total number of users). Consequently, the methodology proposed here establishes that the

---

55 For instance, a job portal might be focused only on a specific segment of the market (e.g. graduate or IT jobs).
selected job portals must have a relatively high number of vacancies, be well-known (traffic ranking) by people, and have a well-defined website structure.

Regarding the former, as shown in Table 5.1, 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. Consequently, these kinds of websites do not necessarily contain a greater number of job advertisements.

Moreover, the amount of information (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 on these websites. In this sense, except for Jobportal_d, the job portals listed in Table 5.1 show similar variables. Indeed, to post a vacancy on these websites, the employer needs to supply a minimum of information (required fields). With some minor variations, this guarantees that these job portals usually have information regarding the job title, city, wages offered, educational requirements, and the company’s name, among others. In contrast, the Jobportal_d website does not have a pre-defined format where employers need to fill in 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.

---

56 Jobportal_d announced that the website had a total of 172,440 job vacancies available on October 30, 2017. However, when clicking on some vacancy announcements, a new window 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 October 30, 2017. However, when clicking on some vacancy announcements, a new window redirected the search and opened another website where the vacancy was originally posted (e.g. Jobportal_a).

57 The magnitude of this issue of redirecting was unknown at this stage of the methodology.
Therefore, considering a job portal like 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 the “trust” of individuals (companies and job seekers) regarding the information provided on a specific website. Additionally, taking into account the traffic ranking of websites guarantees, to some extent, that the selected sites do not specialise in a specific category of vacancies, such as graduate or IT jobs (see Chapter 7 for more evidence regarding this point). As shown in Table 5.1, Jobportal_a, Jobportal_b, and Jobportal_c are the websites with the highest number of visitors.

Table 5.3 summarises the evaluation of job portals conducted in this section. As evidenced in this table, Jobportal_b and Jobportal_a fulfil all the requirements in the consolidation of the vacancy database. These two job portals host the highest number of job advertisements, the variables in the websites are well-defined, and people frequently visit them (traffic ranking).

<table>
<thead>
<tr>
<th>Colombian job portals</th>
<th>Real number of job advertisements</th>
<th>Website quality</th>
<th>Alexa traffic ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.jobportal_a.com.co/">https://www.jobportal_a.com.co/</a></td>
<td>115,723</td>
<td>Well-defined variables</td>
<td>37</td>
</tr>
<tr>
<td>https://www.jobportal_b.com/</td>
<td>62,732</td>
<td>Well-defined variables</td>
<td>89</td>
</tr>
<tr>
<td><a href="https://www.jobportal_c.gov.co/">https://www.jobportal_c.gov.co/</a></td>
<td>Unknown at this stage (this website posts job advertisements that were originally posted on other job portals)</td>
<td>Well-defined variables</td>
<td>199</td>
</tr>
<tr>
<td><a href="https://www.jobportal_d.com.co/">https://www.jobportal_d.com.co/</a></td>
<td>Unknown at this stage (this website posts job advertisements that were originally posted on other job portals)</td>
<td>Not well-defined variables</td>
<td>1,015</td>
</tr>
<tr>
<td><a href="https://www.jobportal_e.com.co/">https://www.jobportal_e.com.co/</a></td>
<td>20,143</td>
<td>Well-defined variables</td>
<td>2,280</td>
</tr>
<tr>
<td><a href="https://www.jobportal_f.com.co/">https://www.jobportal_f.com.co/</a></td>
<td>46,853</td>
<td>Well-defined variables</td>
<td>3,683</td>
</tr>
</tbody>
</table>

*Source: Author’s elaboration.*
As mentioned above, some websites such as Jobportal_c and Jobportal_d (sometimes) redirect the search and open another website where the vacancy has been originally posted (e.g. Jobportal_a). This redirection issue makes it difficult to know the exact number of observations that each job portal can provide to the vacancy database. Given this uncertainty, the other criteria provide more clarity on which portals should be selected. On the one hand, Jobportal_c has a well-defined structure and has a relatively high traffic ranking. Moreover, this portal is a governmental platform, and it might post governmental vacancies that are not available in other job portals. Thus, Jobportal_c should be considered for the vacancy database. On the other hand, as also mentioned above, Jobportal_d does not have a well-defined website structure. Moreover, there is a considerable difference between the traffic ranking of the first three job portals (Jobportal_a, Jobportal_b, and Jobportal_c) and Jobportal_d. Thus, this job portal does not fulfil the criteria to be considered for the consolidation of the vacancy database.

Finally, Jobportal_e and Jobportal_f have a lower number of job advertisements and low traffic ranking. 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 Jobportal_a, Jobportal_b and Jobportal_c websites (Chapter 7 demonstrates that the job portals selected offer a variety of jobs from low-skilled to high-skilled positions).

Consequently, Table 5.4 offers a summary of the web pages that have been selected to be scraped and analysed after an exploration of job portals based on the three elements mentioned above: they have a relatively high number of job announcements (volume), users (traffic), and are well-defined (quality).
Table 5.4. Job portals and their main characteristics

<table>
<thead>
<tr>
<th>Job portal</th>
<th>Main characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobportal_a</td>
<td>It is a widespread private platform in Latin America. 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.</td>
</tr>
<tr>
<td>Jobportal_b</td>
<td>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.</td>
</tr>
<tr>
<td>Jobportal_c</td>
<td>It is a platform administrated by the Colombian Government (more specifically by the 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.</td>
</tr>
</tbody>
</table>

Source: Author’s elaboration.

Finally, it is important to note that the quality and quantity of information provided by these sources might change over time. Moreover, platforms that were not taken into account in this occasion or new ones might start to provide valuable information (increasing the number of advertisements, increasing the number of users, etc.) in the future. This dynamic might change the decision about which job portals to consider for the construction of a future vacancy database. Thus, the 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.

5.4. Web scraping

As was previously seen in Section 5.2, the differences between and within job portals require modifications in programming language and codification structure. Hence, in order 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

---

58 Indeed, there is a version of this platform for Colombia, Peru, Argentina, Uruguay, Guatemala, Ecuador and El Salvador, Honduras, Venezuela, Nicaragua, Cuba and Costa Rica, Mexico, Chile, Panama, Dominican Republic, Bolivia, Paraguay, and Puerto Rico.
the internet (in this case, from vacancy portals) (Oxford Dictionaries 2017). Broadly speaking, this is attained through different softwares that simulate 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 from the data. Codes in R software were built to make this automatic collection of data possible. With the codes developed in this book, the computer can be programmed to visit each job advertisement announcement, to copy all relevant information related to the description of 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 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 languages 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 show the data: the Xpaths change between one website and another. Moreover, so far, there is not an automatic way to determine which Xpaths contain 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 in the HTML structure from one website compared with another, it is necessary to create a different code for each web portal in order to download relevant vacancy information. In consequence, for this book, this method required the construction of three different codes: one for Jobportal_a, one for Jobportal_b, and the third one for Jobportal_c.\(^{60}\)

\(^{59}\) An expression used to identify nodes in websites.

\(^{60}\) 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 other user credentials. This login issue (among other issues)
Once the codes are created, the next step is to run the programs in order to download the corresponding information.\footnote{The process of downloading data using web scraping from a website such as Jobportal\_a can last one day, meaning that the computer visits around 80 announcements per minute to obtain the required information, 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.} Each time the codes are run, the (uncleaned) data are 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.\footnote{For instance, consider a job portal that has 50 vacancies in October 2017, and the corresponding code fails 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, it is critical, first, to periodically review (via visual inspection) that each code 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 document, each code was run three times per month to avoid information loss.

### 5.5. The organisation and homogenisation of information

Once the data are obtained, the next step is to provide a well-defined structure to the semi-structured data collected from vacancy portals. As can be observed makes it necessary to connect R with a software testing framework such as “Selenium” for scraping other websites. Thus, the codes and computing tools (packages and software) to scrape information from job portals might differ between job portals.
in Appendix A, Figure A.4, the localisation (XPath) of a variable might change between job advertisements. 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.

5.5.1. 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” job vacancies, which generates a dummy variable that takes the value of one if a particular pattern can be found in a job advertisement (see Appendix B).

Not all variables can be classified into dummy 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.

5.5.2. 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”. Given the diverse forms that wage information might take, a number of steps were followed. First, an algorithm was programmed that searches and extracts wage information (in whatever form it takes) from job advertisements.

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 Colombian pesos that a worker would receive once hired, no depuration was applied, but where wages were described in words, the words were transformed into their equivalent in numbers. Additionally, when wages were shown in ranges, the average value was selected between the maximum and the minimum range. It is important to note that in the above steps, explicit information about wages was sought, and imputation procedures were not yet implemented (Chapter 6 will discuss the issues regarding missing values and possible ways to handle them).

5.5.3. Company classification

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 as well as on goods and services, companies are classified into sectors. Evidently, the required skills might differ from one industry to another sector. 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 occupations are sector-specific or generic.

Frequently, job portals provide information about the company that is advertising a job position. Part of this information might be useful when identifying

---

63 Around £375-£400 monthly.
64 The usage of the Colombian peso ($) symbol or the word “pesos” (which is the Colombian currency) aided the identification of information regarding wages.
the company’s sector. On the one hand, websites, in some instances, have a pre-defined list of sector categories, so that companies are required to select one category when publishing a vacancy (in some cases more than one category to better describe the company’s activities); however, job portals have their own classification criteria to create a list of sector categories, and information between one job portal and another might not be comparable. Moreover, sector categories used by job portals might be highly aggregated. For instance, Jobportal_b has the category “services.” This option is quite broad and very different types of companies can be classified under this, and therefore the same, sector.

On the other hand, the job description might also give some 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 is an issue as it suggests that the categories or words used by job portals or companies do not adequately describe company sectors for economic analysis. Fortunately, in most cases, alongside vacancy details, job portals also display the business name of the company that has posted the vacancy. Additionally, in Colombia, the Single Business Registry (RUES, for its acronym in Spanish) database is also available.65

Consequently, it is possible to correlate the vacancy and the RUES databases by using company names as a connector between them. However, there are some challenges when trying to merge two databases using company names: misspelling or additional information might exist in either, or both, of the vacancy and the RUES databases. For instance, in the vacancy database, the company’s name might appear as “Exitó,” while in the RUES database the same company might have been registered as “Éxito S.A.” Thus, the names that appear in the vacancy and the RUES databases might be not the same, even when they refer to the same company. This possible difference in names between the two databases might complicate the merging of them. Given this issue, it is necessary to utilise word-based matching methods (better known

---

65 The RUES is a database where people register their companies to pay taxes and receive government benefits. In this database, company names are available along with other relevant information such as their International Standard Industrial Classification of All Economic Activities (ISIC) code.
as “fuzzy merge” methods) to merge two or more databases using words or sentences (in this case, company names). Generally, word-based matching techniques 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 document 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 book are useful to assign an ISIC code to more than half of the 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 Chapter 7 for a more detailed discussion regarding this point).

5.6. Conclusion

Information from 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, processing and analysing information from job portals in a reliable and consistent statistical way is challenging. This chapter has discussed and proposed different solutions to build a robust database of vacancy information from job portals.

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, carrying 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, organising job vacancy information for statistical analysis requires different approaches. Key variables such as wages and required qualifications, among others, are dispersed throughout job advertisements. Thus, it is necessary to program an algorithm that deals with linguistic issues (such as gendered words in Spanish), reads each job announcement, and creates an indicator variable that takes a value (for example, 1) if a particular pattern emerges on a job advertisement. However, in order to build a variable such as company 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 in Colombia.

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 chapter will describe the methods that can deduce skill and occupational information, among other relevant variables.