6. Extracting More Value from Job Vacancy Information (Methodology Part 2)

Jeisson Cárdenas Rubio
6. Extracting More Value from Job Vacancy Information
   (Methodology Part 2)
6.1. Introduction

The previous chapter has shown that information from job portals 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 Chapter 2, 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. (OECD 2016a; 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 Chapters 2 and 3). 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 collecting detailed labour demand information (Chapter 4). Thus, the key task of this chapter is to describe the techniques that can be utilised to extract information on skills and occupations.

As mentioned in Chapter 5, 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 job titles, respectively. Consequently, this chapter explains the steps required to organise and categorise skills and occupational information from the vacancy database. Section 6.2 of this chapter develops a methodology to identify skill patterns in job vacancy descriptions based on international skill descriptors, such as the ESCO. However, there might be some country-specific skills that are not listed in the ESCO dictionary, or its international skills descriptors might not be updated according to the most current labour demand
requirements. Therefore, Section 6.3 proposes a methodology to automatically identify country-specific or new skills from information from job portals.

The classification of job titles into occupations is a critical stage for vacancy analysis. Correctly coding the job title variable requires different and advanced data mining techniques. Therefore, Section 6.4 describes and applies techniques such as manual classification, software classifiers, and machine learning to organise job titles into occupational groups. This section also proposes a method that uses unstructured information from job titles and skill 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., this 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, Section 6.5 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 Chapter 5, 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, Section 6.6 outlines how missing values were imputed for the “educational requirement” and “wage offered” variables by using predictors such as occupation, city, and experience requirements, among others. Finally, Section 6.7 presents consolidated, organised, categorised, cleaned, and imputed data for the analysis of the Colombian labour demand using job portal sources. Next, Figure 6.1 provides a summary of the above described steps that were implemented to organise Colombian vacancy information.
6.2. Identifying skills

As shown in Chapter 5, 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 Chapter 2 for more discussion on the skill concept). For illustrative purposes, Table 6.1 shows an example of a job description.
As highlighted in Table 6.1, some words or phrases in the job description can be associated with the skill concept. More specifically, words such as “office automation” (“ofimática” in Spanish) or “environmental management” (“gestión ambiental”) can be seen as specific skills required for this vacancy.

Unlike for the study carried out by 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 separate 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 Chapter 2, there are different ways (typologies or dictionaries) to organise and analyse information regarding skills. Consequently, the first step to organise this kind of information dispersed within vacancy advertisements is to select a dictionary of words or phrases related to skills. Through this

---

Table 6.1. **Job description**

| Description
|---|

**Source:** Jobportal_a.
method, it is possible to identify patterns (words or phrases) that are connected to skills in 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 the O*NET and the ESCO.

As mentioned in Chapter 2, the O*NET is based on the US Standard Occupational Classification (SOC) system. This system contains information on hundreds of standardised and occupation-specific descriptors. 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.

The 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 disaggregation for each occupation, such as an exhaustive list of 13,485 relevant skills (skills pillar) (European Commission 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 the O*NET: since the ESCO is mapped following the ISCO-08 structure, the two systems of classification (ESCO and ISCO-08) are compatible. As the ESCO handbook points out: “This is particularly important because most national occupational classifications are currently mapped to ISCO-08” (European Commission 2017, p. 29). Indeed, in 2015, Colombia accepted recommendations made by the International Labour Organization (ILO) to adopt ISCO-08 as official classification.67 Thus, to obtain results compatible with the official national classification for this book, 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. First, common words in the Spanish language (such as

---

prepositions, stop words) were removed from the ESCO dictionary and from job descriptions 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 descriptions 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-grams), which are indicator variables. Variables take the value of one if a certain a word or phrase (pattern) of the skills dictionary is found in an advertisement, and zero 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 required skills. Consequently, as will be discussed in more detail in the next chapter, unigram variables show the key skills required 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 job descriptions helps to identify the key skills in demand within the Colombian labour market. Additionally, as shown in Section 6.3, 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 Section 6.4). Moreover, as will be seen in Subsection 6.4.7, unigram skill variables facilitate assigning occupational codes to the vacancy database.

6.3. Identifying 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 that 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 Chapter 2, 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 skills dictionaries might cause that these dictionaries do not adequately measure what skills are currently 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 dictionary were removed from the description of each vacancy. Nevertheless, several words remained that did not correspond with 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 Section 6.2. 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.

It is necessary to note that the words that remain in a job description after applying this method might still contain terms that are not related to skills. For instance, there might appear words related to places or names of people, companies, etc. Moreover, words might appear that refer to other characteristics of the potential worker, such as physical attributes. Consequently, based on the skills definition of this book (see Chapter 2), the final step consists of a visual and manual inspection of the words that remain in the job description to determine which of them are describing new and/or specific skills (Chapter 7 will show a list of new and/or specific skills demanded by the Colombian labour market).
6.4. Classifying vacancies into occupations

One of the most critical variables is “job title” because it summarises the main characteristics of the labour demand and allows classifying jobs into occupations (or skills). Figure 6.2 presents a frequency analysis of the words that appeared in job titles in Colombia in January 2017. According to this figure, the most frequent words appearing 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”), and Drivers (“Conductor”), among others.

Figure 6.2. Word cloud: Frequency analysis

Source: Author’s elaboration based on vacancy database, 2017.

The text mining figures are presented in the Spanish language because it is the original language used in the Colombian job portals.
Figure 6.3 shows in more detail the words that are most associated with job titles. 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 6.3, 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” (“Ase- sor”) group, the most frequent occupations are in “Sales, commerce, and customer service.”

Figure 6.3. Word association: Frequency analysis

Source: Author’s elaboration based on vacancy database, 2017.

The above figures are approximations 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 6.2 and Figure 6.3, “Assistant” and “Auxiliar” are considered as different categories, even though they can, on many occasions, refer to the same job category. To avoid these issues 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” (Salazar-Xirinachs 2017).
Regarding job titles, this research seeks to classify all the information available to ISCO-08.\textsuperscript{69} 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 will 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 collect information on occupational titles. Nevertheless, in the case of vacancy data collected from the internet, classifying job titles into occupations might be even 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 the job title information. Moreover, as mentioned above, companies might use a variety of different words to describe the same occupation. This issue makes the classification of job titles into occupational codes a challenge.

Given the complexity of classifying job titles into occupations and the importance of this information for researchers, the 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, as Gweon et al. (2017) note, 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 the Colombian vacancy data\textsuperscript{70}. 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.

\textsuperscript{69} As previously mentioned, Colombia accepted the recommendations made by the ILO to adopt ISCO-08 (ILO 2008) as an official classification for jobs.

\textsuperscript{70} For instance, the Colombian vacancy data collected for this document 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.
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 softwares 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” (using ISCO-08). The latter term—completely automated coding—refers to methods such as machine learning. Briefly, these sets 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 exists, 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, there exist softwares such as Cascot (Computer Assisted Structured Coding Tool71) (Jones and Elias 2004) (see Subsection 6.4.3) that allow 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 book 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).72

---

71 Developed by the Institute for Employment Research (IER) at the University of Warwick.

72 To avoid misclassifications, Jones and Elias (2004) recommend the implementation of both partial and fully automated coding (semi-automatic coding). For instance, in the case of Cascot, 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.
Thus, given the availability of several tools to classify occupations and the advantages and disadvantages of each one of them, the next sections will discuss manual coding, cleaning, automatisation, and adapting Cascot.

6.4.1. Manual coding

As pointed out before, manual coding is a time-consuming task. However, as shown in Figure 6.2 and Figure 6.3, there are some job titles that are more frequently mentioned by employers, hence those job positions constitute a considerable share of the vacancy database. Additionally, automatic algorithms might misclassify some job titles, given that automatic methods of classification might fail to classify 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 manual coding process was carried out for job positions that were more numerous, and, therefore, it was relatively easy to determine their occupational groups. 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 plural. 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 the Colombian vacancy information is concentrated across relatively few job titles. 73

6.4.2. Cleaning

As mentioned above, coding quality depends on the tool used and on the quality of input data. However, job titles displayed on job portals sometimes contain extra information (noise) that might affect coding quality. While there

73 At this point, this result neither validates nor invalidates the reliability of data. The Colombian labour market might demand a particular set of occupations (see Chapter 7 for further discussion).
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 6.2, 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. In order 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” with a similarity index of 92, Cascot assigns the ISCO code 2411 (“Accountants”)—in a scale of 0 to 100, the higher the number, the higher the degree of certainty that a given code is the correct one. However, when the input is “Accountants immediately,” the similarity index drops to 66.

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 data. Second, the variables “city” and “company name” (provided that the structure of the website contained this information) were used to identify all possible locations and employer names that might arise in job title variables. With this process, names that might appear in the job title were dropped. 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 cleaning process, automatic classification tools and techniques were used.

6.4.3. 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
A web-based approach to measure skill mismatches and skills profiles for a developing country

(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. (Warwick Institute for Employment Research 2018)—that 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.org 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 deep 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 does not exist (or at least is not publicly available) a trustworthy pre-processed database with job titles and occupational codes. Machine learning methods need as input a training database (which is data that were 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
when the similarity score was greater than 45. This threshold was to re-ensure that Cascot outputs would coincide with the manual coding revision in most cases. By doing so, around 38% of the observations in the vacancy database received an occupational code at the four-digit level.\textsuperscript{74} Thus, 35\% of the job advertisements required further data management to assign a proper occupational code.

6.4.4. 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 explained in Subsection 6.4.1 was applied. First, a visual inspection of the vacancy database was conducted on data that were 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 issue with gendered or plural (singular) forms. This 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 the observations were assigned an occupational code with a relatively high standard level of confidence.

6.4.5. Adaptation of Cascot 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 the ILO (2008, p. 68): “[occupational titles provided by ILO] might be a good starting point to develop a national index. The national index, 

\textsuperscript{74} 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 the observations had the correct occupational code (ISCO-08) at a four-digit level. Moreover, common mistakes were manually corrected.
however, needs to reflect language as used in survey responses in the country concerned.” Even in countries with the same language, job positions might be named differently depending on the national context. Consequently, standard occupational titles provided by the ILO might not cover a considerable share of Colombian job titles, hence Cascot might not assign an occupational code to a high portion of them. Indeed, this issue of context might explain that, at this point, only 38% of the job portal observations were categorised using Cascot.

Moreover, the DANE released an adaptation of the ISCO occupational titles according to the Colombian context in 2015 (DANE 2015). Thus, given that Cascot can be edited, 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 ready, the job titles that had not been 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 6.4.3. Interestingly, with the adaptation of the tool, around 12% of the total number of advertisements were assigned an occupational code. Thus, by only adapting the Cascot tool using the national occupational titles of Colombia, the portion of job advertisements has considerably increased 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 the joint work of institutions such as the 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.  

75 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 like Peru or Chile (where Spanish is also the official and most spoken language).

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

As a result of the above described 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 versions. By doing this, 3% of the job titles in the vacancy database received an occupational code.

At this point, 15% of the 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 explained repeatedly above, is a time-consuming task. Therefore, this option was not considered. At the same time, the second and third options contain various advantages and disadvantages. On one hand, the Cascot similarity threshold could be lowered to classify more job titles (so far, the threshold has been 45). Nevertheless, this might increase the number of misclassified observations. 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 research year to assign occupational codes to job titles. Depending on the sophistication of their algorithms and

---

77 This option is the most straightforward alternative to assign 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 the job titles were incorrectly classified. Thus, lowering the Cascot threshold was not an option to classify the remaining job titles.
inputs (training and test databases), these techniques might adequately assign occupational codes to job titles (Bethmann et al. 2014).

### 6.4.7. Machine learning

The use of machine models that classify job titles into occupation codes has arisen over the last decades. As Gweon et al. (2017) highlight, 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). Moreover, as Murphy (2012) points out, classification (Supervised Learning) 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, 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 D). For the case of job titles, 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 (Support Vector Machines) (Gweon et al. 2017) (see Appendix E). However, as Appendix F demonstrates, 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.

---

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

79 Unsupervised and reinforcement learning are other types of machine learning algorithms. However, as the purpose of this subsection is the classification of job titles, this document is focused on Supervised Learning.
6.4.7.1. 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 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-grams 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).

6.4.7.2. Machine learning using skills

Conversely, Lima and Bakhshi (2018) proposed an extension of the basic machine learning model for classifying job titles into occupations. For this study, the authors used UK job vacancies published in 2015, collected by Burning Glass Technologies.80 This company assigns each vacancy one or more of 9,996 tags derived directly from the job advertisement text (the authors did not clarify, though, 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 doing so, Lima and Bakhshi (2018) demonstrated that a skills-based classifier might improve the coding of jobs titles that are poorly classified.

---

80 Burning Glass Technologies is a company that provides job market analytics.
6.4.7.3. Nearest neighbour algorithm using skills and job titles

Given these advantages and limitations of the more recently proposed algorithms, this document uses an extension of the algorithm proposed by Gweon et al. (2017) 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 the 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).

6.4.7.3.1. Application of the extended nearest neighbour algorithm to the vacancy database

As mentioned in Subsection 6.4.6, 15% of the job titles remained uncoded at this stage through manual and Cascot procedures. Consequently, the final step to classify the remaining job titles was conducted using the extended nearest neighbour algorithm, explained in Appendix G (Tables G.5 and G.6). However, as pointed out in Section 6.2, unlike the study of 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 book uses the n-gram skill variables created in Section 6.2 as an input for the algorithm proposed here.

Specifically, the 1,910,000 observations (85% of the vacancy database) coded from Subsection 6.4.1 to 6.4.6 were used as input to train and test the extended nearest neighbour algorithm. Each of those observations has the corresponding 4-digit-level ISCO codes and the n-gram skill variables identified in this book. Moreover, this input database was divided into two: training and test database. Following Dobbin and Simon (2011), the training dataset
is composed by 1,273,333 (two-third) observations (randomly assigned) from the input database, while the test database is composed of the remaining one-third of the input data. The computer determines how to classify the job titles by executing the extended nearest neighbour algorithm with the training database. Once the computer had learnt the association rules, the algorithm was executed in the test database. The predicted results were compared with real ISCO codes in the test dataset. The comparison showed that the extended nearest neighbour algorithm correctly classified 92% of the test dataset. Thus, the algorithm showed a high accuracy prediction level.

By doing so, this book 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 (see Appendix G for a comparison between the accuracy level of these algorithms). With this method, 10% of the 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 of writing this book, it is still necessary to use manual coding for those job titles that 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).81

81 Importantly, a considerable 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 kinds 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 in 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 part of a future work.
6.5. Deduplication

Along with the categorisation challenges shown above, there is another important issue to consider, which is the possibility of duplicated information. As data are collected from different websites, some job advertisements can appear on more than one job board, or even several times on the same job board (Chapter 4). This issue can result in a significant over-counting of job advertisements and might affect the results when data are 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, Jayasundera, and Repnikov 2014).

One option is to drop those vacancies that 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 methods for dropping duplication in job titles are preferable than string-based methods. As mentioned in Section 6.2, n-grams are a set of indicator variables based on text patterns. The variables take the value of one in the presence of specific patterns.

Consequently, n-gram-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 the previously created dummy variables (such as experience, educational requirements, type of contract, localisation, and wages) were the same as in other job advertisements, including their ISIC (Chapter 5) and ISCO codes (Section 6.4), the publication date, and the number of job positions required. By doing so, around 26% of the observations were discarded.
6.6. Imputing missing values

Provided that the information comes from websites and employers who might not provide a full description of the vacancy, there are variables with missing values. For instance, despite the text mining techniques explored in Chapter 5, around 30% of the 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 full data vacancy information.

In this regard, Carnevale, Jayasundera, and Repnikov (2014), with hot-deck and cold-deck methods, imputed missing educational requirements in job advertisement data using a combination of the education distribution of the vacancy data (no missing values) 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 whether 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 document, an imputation procedure is conducted for the wage and educational variables.

6.6.1. Imputing educational requirements

For the Colombian case, 20% of the 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 missing values, a hot-deck imputation was conducted as proposed by Carnevale, Jayasundera, and Repnikov (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 requirements” receives a value from an observation that is randomly selected from a sample of records, which have the same characteristics in common, such as the same occupation. Consequently, as a first step, it is necessary to define what characteristics define the sample of donors (“deck”) for an observation with a missing value.

Within a vacancy, the variables occupation, city, and year were considered as characteristics that 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 about 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 important 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 for 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 (2006) 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, it is necessary to have a higher educational level given technological changes, among other factors. Besides these, 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 requirements” 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 book did not implement 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 book does not 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, Jayasundera, and Repnikov (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 by Carne-
vale, Jayasundera, and Repnikov (2014) uses the American Community Survey
(ACS) (which is a survey on labour supply) to impute missing values in the
job vacancy data. However, as will be discussed in more detail in Chapter 7,
missing vacancy values based on a household (supply) survey might be problem-
atic due to the distribution of educational requirements (among other character-
istics) that might differ between labour demand and labour supply. Moreover,
part of this book seeks to test whether the vacancy database shows consistent
patterns compared to 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 this cold-deck imputation process.

6.6.2. Imputing the 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 can be applied, such as the LASSO regression
(“Least Absolute Shrinkage and Selection Operator”). Unlike linear models,
the LASSO model penalises 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-ma-
jor 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:

\[
y = \beta_1 \text{Occupation}_i \chi_{i=1\ldots40} + \beta_2 \text{Department}_i \chi_{i=1\ldots32} + \beta_3 \text{Quarter}_i \chi_{i=1\ldots4} + \beta_4 \text{Education}_i \chi_{i=1\ldots8} + \beta_5 \text{Workday}_i \chi_{i=1\ldots3} + \beta_6 \text{TypeContract}_i \chi_{i=1\ldots4} + \varepsilon
\]

Where \( y \) is the wage variable, “Occupation” denotes the set of dummy variables that identify occupation (ISCO two-digit level, 33 subgroups);\(^{82} \) “Department” represents the set of dummy variables that identify the department where the vacancy is available (there are 32 departments 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 that indicate educational requirements (six categories\(^ {83} \)); “Workday” and “TypeContract” are sets of dummies variables indicating the workday (three categories) and the type of contract (four categories) offered by employers (all of these categories will be explained in more detail in Table 6.2).\(^ {84} \)

### 6.7. Vacancy data structure

Figure 6.4 summarises the steps carried out and the amount of information processed to consolidate the vacancy database for Colombia.

---

\(^{82}\) The occupation variable was grouped at a two-digit level to avoid oversaturation and due to computational limitations.

\(^{83}\) Due to frequency issues, the categories of specialisation, master’s degree, and doctoral degree were grouped in one category: “Postgraduate.”

\(^{84}\) The variable sector was not included in the imputation model due to the high frequency of missing data.
Based on these steps, this book provides a robust methodology to process and organise information from job portals. As a result, the Colombian vacancy database created this way has the following structure as detailed in Table 6.2. \(^{85}\)
Table 6.2. **Basic data structure**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Percentage of missing values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job title</td>
<td>Short description about the job title</td>
<td>No missing values (mandatory field in the job advertisement)</td>
</tr>
<tr>
<td>Vacancy description</td>
<td>Detailed information about the profile required to fill the vacancy</td>
<td>No missing values (mandatory field in the job advertisement)</td>
</tr>
<tr>
<td>Labour experience</td>
<td>Dummy variable that takes the value of 1 if the vacancy (explicitly) requires any labour experience and 0 otherwise</td>
<td>No missing values (this variable takes the value of 0 if a vacancy does not say anything related to labour experience)</td>
</tr>
<tr>
<td>Number of vacancies</td>
<td>Number of job positions offered for each job advertisement</td>
<td>No missing values (mandatory field in the job advertisement)</td>
</tr>
<tr>
<td>Company name</td>
<td>Name of the company who published the job advertisement</td>
<td>Around 4.5% of job advertisements with missing values</td>
</tr>
<tr>
<td>Publication date</td>
<td>Starting date when the job advertisement was placed</td>
<td>Around 20.0% of job advertisements with missing values</td>
</tr>
<tr>
<td>Expiration date</td>
<td>Date when the job advertisement expires</td>
<td>Around 65.3% of job advertisements with missing values</td>
</tr>
<tr>
<td>Educational</td>
<td>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’s degree; h. doctoral degree. See Chapter 8.</td>
<td>Around 20.0% of job advertisements with missing values. After the imputation process, no observations had missing values in this variable.</td>
</tr>
<tr>
<td>Educational</td>
<td></td>
<td>No missing values (this variable takes the value of 0 if a vacancy does not say anything related to education)</td>
</tr>
<tr>
<td>Wage</td>
<td>Continuous variable that indicates the amount of money that the hired person will receive</td>
<td>Around 30.0% of job advertisements with missing values. After the imputation process, no observations had missing values in this variable.</td>
</tr>
<tr>
<td>Imputed wage</td>
<td>Continuous variable that indicates the amount of money (imputed) that the hired person will receive</td>
<td>No missing values</td>
</tr>
<tr>
<td>Type of contract</td>
<td>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</td>
<td>No missing values (this variable takes the value of 0 if a vacancy does not say anything related to type of contract)</td>
</tr>
<tr>
<td>Workday</td>
<td>Set of dummy variables that identify the workday offered by the employer: a. full-time; b. part-time; c. by hours</td>
<td>No missing values (this variable takes the value of 0 if a vacancy does not say anything related to workday)</td>
</tr>
<tr>
<td>City</td>
<td>Place where the vacancy is available</td>
<td>Around 1.2% of job advertisements with missing values</td>
</tr>
<tr>
<td>Sector ISIC</td>
<td>ISIC Code (2 digits if possible)</td>
<td>Around 39.1% of job advertisements with missing values</td>
</tr>
<tr>
<td>Variable</td>
<td>Definition</td>
<td>Percentage of missing values</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Skills</td>
<td>Set of dummy variables that identify the skills required by employers according to ESCO</td>
<td>No missing values (this variable takes the value of 0 if a vacancy does not say anything related to skills)</td>
</tr>
<tr>
<td>Specific skills</td>
<td>Set of dummy variables that identify (country-specific) skills required by employers and are not listed in the ESCO dictionary</td>
<td>No missing values (this variable takes the value of 0 if a vacancy does not say anything related to specific skills)</td>
</tr>
<tr>
<td>ISCO Code</td>
<td>ISCO Code (4 digits if possible)</td>
<td>Around 4.2% of job advertisements with missing values</td>
</tr>
</tbody>
</table>

Source: Author’s elaboration.

6.8. 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 occupations required by employers. Nevertheless, to obtain consistent information for skills and occupational requirements from job advertisements, it is necessary the use of dictionaries or classifications, along with the implementation of more complex algorithms. Consequently, the first part of this chapter 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 chapter showed that the usage of international dictionaries such as the ESCO might facilitate building a methodology that identifies the skills demanded in each job advertisement for countries like Colombia. Moreover, with the help of text mining techniques, it 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, potentially, be 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 it 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 its 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 chapter, 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 does not exist a unique method that satisfactorily assigns occupational codes to job titles. Given the advantages and disadvantages of each approach, this document 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 neighbour 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 are collected from different websites (some job advertisements can appear on more than one job board or even on the same job board), the second part of this chapter showed how to deal with duplicated records. Specifically, it was argued that an 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 no 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 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 chapter explained and used the hot-deck and LASSO methods to impute missing values into the “education required” and “wage” variables.

In summary, this chapter 1) provided a robust and detailed methodology to obtain, organise, and categorise skills and occupations from job portals for statistical analysis; and 2) showed how to deal with duplicated job advertisements and missing values for relevant variables. Thus, as an outcome of this and the previous chapter (Chapter 5), the vacancy database can now be tested.