1. Introduction

Jeisson Cárdenas Rubio
1. Introduction
This book studies how, and to what extent, a web-based system to monitor skills and skill mismatches could be developed for Colombia based on information from job portals. More specifically, this document seeks to answer the following questions: 1) How can information from job portals be used to inform policy recommendations? And, in order to address two of the major labour market problems in Colombia, which are high unemployment and informality rates, 2) to what extent can information from job portals (unsatisfied demand) and national household surveys (labour supply) be used together to provide insights about skill mismatch issues in a developing economy?

Consequently, this book investigates the challenges, advantages, and limitations of collecting information from job portals and proposes a framework to test this information’s validity for economic analysis. It conducts an innovative labour market analysis and develops indicators based on updated and robust labour demand (job portal) and labour supply (household survey) information to tackle skill mismatches, extending thus the use of novel sources of information to yet unexplored areas in the existing labour economics literature.

By doing so, this study makes conceptual, methodological, and empirical contributions to the ongoing debate in economics about the use of information from job portals for labour demand analysis. The main conceptual contribution consists of demonstrating that the concept and sources of Big Data (in this case, job portal sources) can provide consistent results to orient public policies (see Chapters 7 to 9). This document also demonstrates that, with the proper techniques, information from job portals can fulfil conceptual requirements to be considered as high-quality data for labour market analysis (see Chapters 4 and 10).

The main methodological contribution is the development of a detailed framework and methods to collect, clean, and organise (i.e. web scraping, occupation and skill identification, etc.) vacancy data, which allows testing and analysing this source of information for consistent labour market insights.
Specifically, this book contributes to the methodology of processing information from job portals for public policy advise by: 1) discussing different criteria (volume, website quality, and traffic ranking) to select the most relevant and trustworthy job portals in order to collect vacancy information (Chapter 5); 2) providing a detailed explanation about Big Data techniques (web scraping) and the challenges they pose for automatically collecting job advertisements from job portals (Chapter 5); 3) applying mixed-methods approaches (text mining, word-based matching methods, etc.) to standardise information collected from different job portals into a single database for statistical analysis (Chapter 6); 4) implementing and extending a mixed-methods approach (stop words, stemming, extensions of a machine learning algorithm, etc.) in order to identify skills and occupations in online job announcements (Chapter 6); 5) and, importantly, using this extended mixed-methods approach (e.g. a skills dictionary to identify skill patterns) to find new or specific skills and occupations in the Colombian labour market, which would otherwise be complex to identify via other means (e.g. household surveys) (Chapter 6).

Moreover, the book proposes a (n-gram-based) method to reduce duplication issues (as information is collected from different job portals, some job advertisements can be repeated) and a (Lasso) method to impute missing values, such as education and wages (Chapter 6). Consequently, by implementing and extending novel mixed methods, 6) this document improves data collection and helps to understand methodological changes to collect and organise information from job portals.

As a product of the above methods, a vacancy database was consolidated for the period between January 1, 2016 and December 31, 2018 (Chapter 7). In addition, this document makes further methodological contributions by 7) proposing a framework to evaluate the internal (consistency) and external (representativeness) validity of this vacancy database. To test internal validity, a statistical comparison was conducted between variables, such as wages, occupations, education, etc., to understand biases, errors, and inconsistencies within the database. The evaluation of external validity was particularly challenging because countries like Colombia do not have vacancy censuses (or anything similar) to compare information collected from job portals. Despite several obstacles, this book provides and applies a methodology framework to evaluate the vacancy database. It implements a detailed comparison between official
information available in the country (i.e. household surveys) and vacancy data results, such as vacancy, employment, new hires, unemployment, occupational structures and their dynamics over the study period. This comparison enables the understanding of possible biases (e.g. over/underrepresentation of certain occupational groups) in the vacancy database (Chapter 8).

Based on the validation results, another methodological contribution of this document is (8) proposing and estimating skill mismatch measures that consider the advantages and limitations of job portals and household surveys. Specifically, the study demonstrates how household surveys can be combined with vacancy data to produce relevant (volume- and price-based) skill shortage indicators, such as percentage change in unemployment by sought occupation, percentage change in median real hourly wage, among others. Importantly, (9) this book makes an important contribution to the discussion about skill mismatch measures by considering informality. As will be discussed in Chapter 9, informality is a signal of labour market imbalance. A considerable portion of employment growth might be explained because people cannot find a formal job and have to choose informal jobs. Thus, skill shortage indicators need to control for informality to avoid misleading results.

Based on the above methodology, this book also makes relevant empirical contributions by providing a detailed labour market analysis that reveals important characteristics of the Colombian labour demand (e.g. demanded skills and occupational trends). Importantly, it determines skill mismatches (i.e. skill shortages) in Colombia based on information from job portals and household surveys. Specifically, the analysis of the vacancy database evidences that (1) data collected from job portals are representative of a considerable set of non-agricultural, non-governmental, non-military, and non-self-employed (“business owners”) occupations; 2) most of the vacancies in Colombia correspond to middle- and low-skilled occupations (such as “Sales demonstrators”); 3) in alignment with the most demanded occupations, the most demanded skills are “Customer service,” “Work in teams,” etc.; and, most importantly, 4) information from job portals can be used to identify new or specific job titles (e.g. “TAT vendors,” “Picking and packing assistants,” etc.) and skills (e.g. “Siigo,” “Perifoneos,” etc.) for the Colombian context.

Based on the advances made towards homologating vacancy and household survey information (e.g. coding both databases according to ISCO-08),
A comprehensive analysis of labour demand and supply information is conducted at the occupational level (Chapter 9), for the first time in Colombia. Another important contribution of this analysis consists of 5) showing in detail population groups with higher (lower) informality and unemployment rates. For instance, domestic cleaners and helpers and motorcycle drivers face the highest informality, while environmental engineers and geologists and geophysicists face the highest unemployment rate in the country. In addition, 6) it also estimates skill shortages using job portals and vacancy information. For instance, it evidences that 30 occupations show signals of skill mismatches, while indicating that Structured Query Language (SQL), database management, and JavaScript are the most demanded skills for one of those occupation groups (“Web and multimedia developers”).

Briefly, skill mismatches arise when there is a misalignment between the demand and supply of skills in the labour market (UKCES 2014). As will be discussed in Chapters 2 and 3, numerous multidisciplinary studies have pointed out the importance of these phenomena in labour market outcomes, such as unemployment and informality, among others. Skill mismatches can occur in the job search process (e.g. skill shortages) or in the workplace (e.g. skill gaps). Given that the term “skill mismatches” encompasses different dimensions and considering available data to analyse an economy such as Colombia (i.e. job portals and household surveys), this book focuses on studying skill shortages. This concept refers to issues that arise in the job searching process when jobseekers do not have the proper skills required in vacancies posted by employers (Green, Machin, and Wilkinson 1998).

A proper labour market analysis system to identify possible skill shortages and current employer skill requirements is paramount for a country such as Colombia with high and persistent unemployment and informality rates (DANE 2017a). According to the Colombian statistics office (National Administrative Department of Statistics; DANE for its acronym in Spanish), in the last two decades unemployment and informality rates were around 12.5% and 49.4%, respectively. A vast number of factors, such as rigid wages, comparatively high non-wage costs, etc., could explain these labour market outcomes. However, as will be discussed in Chapters 2 and 3, theoretical and empirical evidence shows that mismatches between demanded skills and those offered is a main cause of unemployment and increased informality rates in Colombia (Álvarez and
Hofstetter 2014; ManpowerGroup, n.d.; Arango and Hamann 2013). Workers, the government, as well as education and training providers are not properly anticipating employer requirements. Consequently, the labour supply lacks skills in relation to what employers are demanding in order to fill their vacancies.

Despite evidence that suggests that there is a high incidence of skill shortages in the Colombian labour market, education and training providers, workers, and the government can do little to reduce imperfect information regarding human capital requirements due to a lack of proper information to develop well-orientated decisions and public policies (González-Velosa and Rosas-Shady 2016). On the one hand, the cost of conducting household or sectoral surveys (traditional sources of information) is relatively high in terms of resources and time. On the other hand, these data sources usually fail to provide detailed and updated information about skills and occupational requirements. These issues have discouraged countries (especially those with low budgets) from collecting information on and analysing human capital needs.

For instance, the Colombian office for national statistics (DANE) periodically conducts household and sectoral surveys that provide valuable insights about the characteristics of the Colombian workforce, job training, selection and hiring practices, productivity, etc. However, due to sample constraints and the relatively high operational cost of conducting these surveys (e.g. the job of interviewers and statisticians, etc.), the data collected do not convey detailed information about employer requirements—the occupational structure demanded—nor about the skills required for each position. Thus, the characteristics and dynamics of labour demand remain relatively unknown.

Consequently, to fill these critical information gaps, it is vital to seek new ways of analysing labour demand that can consistently complement existing surveys (e.g. household surveys). Big Data have become a trendy field because it deals with the analysis of large data sets, in real time, from different sources of information (Edelman 2012; Reimsbach-Kounatze 2015). Using job portals and Big Data techniques to analyse employer requirements constitutes an alternative that has attracted the attention of researchers and policymakers. Employers post a considerable number of vacancies on online job portals along with detailed candidate requirements (job title, wages, skills, education, experience, etc.), which provides quick access to a large amount of relevant information for the analysis of labour demand. This online data can provide key
insights about labour demand that previously were not accessible for proper analysis (Kureková, Beblavy, and Thum 2014).

Collecting, processing, and analysing information from job portals through reliable and consistent statistical processes is challenging because data are dispersed across different websites and the information is not categorised or standardised for economic analysis. Additionally, the discussion regarding the use of Big Data sources, such as job portals for labour market analysis, is flawed (Kureková, Beblavy, and Thum 2014). Different authors have used and derived conclusions from job portal data without considering in detail the possible biases and limitations of this information (e.g. Backhaus 2004; Kureková, Beblavy, and Thum 2016; Kennan et al. 2008). Like any other source of data, information from job portals has biases and limitations. For instance, given the type of internet users, among other data quality issues, job portals are unlikely to be representative of the whole economy or a specific sector, or they might not reflect real trends in labour demand. The lack of debate concerning data validity has affected the credibility of job portals as a consistent and useful resource for labour market analysis.

A conceptual and methodological framework is required in order to use vacancy data and to properly address issues such as skill mismatches. Therefore, this book seeks a better understanding about the use of new sources such as job portals to analyse the labour market (skill mismatches) in a developing country such as Colombia. This study responds to the need to develop a more efficient way to collect and analyse information about labour demand and skills in order to identify potential skill shortages. This kind of work supports the design of national skills strategies, while enhancing the capacity of governments to develop public policies to tackle current skill mismatches (Cedefop 2012a).

To this end, this book is structured as follows: Chapter 2 discusses the concepts and theoretical framework used in this document to analyse labour market based on the information found on online job portals. First, this chapter introduces basic conceptual and statistical definitions for labour demand (e.g. job vacancies) and labour supply (e.g. unemployed and employed workers). Second, given that a considerable share of the population in Colombia works in irregular market conditions, this chapter discusses what is understood in the academic literature by informality. Furthermore, the concept of skills and different ways to measure them for economic analysis are examined. Subsequently,
the previously mentioned definitions are used to describe the dynamics of the labour market and its main outcomes, such as unemployment, wages, etc., under the assumption of perfect competition (e.g. assuming that companies and workers are perfectly informed about the quality and the price of “labour”). Nevertheless, the assumptions of perfect competition are unrealistic given that workers are usually not perfectly aware of employer skill requirements; similarly, this model is not an appropriate theoretical framework for economies such as Colombia (Garibaldi 2006). Based on a model with imperfect information (which seems more appropriate to describe Colombian labour market outcomes), Chapter 2 explains how skill mismatches can arise, as well as their consequences for informality and unemployment rates (Bosworth, Dawkins, and Stromback 1996; Reich, Gordon, and Edwards 1973; Stiglitz et al. 2013). This framework highlights that information failures might be one of the leading causes of high unemployment and informality rates. Thus, actions to decrease these information failures (such as the use of job portals) will considerably improve people’s employability.

Chapter 3 presents evidence that skill shortages, unemployment, and informality are high-frequency phenomena in Colombia (DANE 2017a; ManpowerGroup, n.d.; Arango and Hamann 2013). Moreover, it outlines how the government, as well as education and training providers, etc., face severe difficulties to tackle these issues due to the lack of a proper system to identify skills in demand and possible skill shortages (González-Velosa and Rosas-Shady 2016). First, the chapter describes the main characteristics of the Colombian labour market, such as unemployment, informality, etc., and their evolution during the last two decades. In addition, it provides a general description of the socio-economic characteristics of the labour force and—based on the little information available—the labour demand. Second, it evidences a high incidence of skill shortages in Colombia and their possible implications for labour market outcomes. It is argued that workers, education and training providers, as well as the government can do little to address these issues given the lack of proper information to monitor and identify employer requirements and possible skill shortages at the occupational level. Subsequently, the chapter presents an overview of the Colombian labour market focused on unemployment, informality, and skill shortages, and highlights the need for detailed information to adequately address these issues.
In Chapter 4, the concept of Big Data is introduced, with its advantages and limitations outlined for a labour market analysis. Moreover, this chapter explains why traditional statistical methods, such as household or sectoral surveys, encounter difficulties in providing detailed information about the labour market. First, it defines Big Data according to three properties: volume, variety, and velocity (Laney 2001). Then, it discusses the problems of traditional statistical methods, such as sample or survey design, that constrain labour market analysis in terms of occupations and skills (Kureková, Beblavy, and Thum 2014; Reimsbach-Kounatze 2015). Given these information gaps, the potential use of Big Data sources to complement labour market analysis is discussed, with a special focus on job portals and their possible application to tackle skill shortages. Subsequently, this chapter explains the limitations and caveats to be considered when online vacancy data are used for economic analysis. Furthermore, it emphasises the differentiating features of this book, compared to other ongoing studies.

Once the conceptual framework and the need for information and analysis to address skill shortages are established, Chapters 5 and 6 present a comprehensive methodology to systematically collect and standardise vacancy information from job portals. Chapter 5 describes available information that can be collected from Colombian job portals. Then, it proposes criteria to consider the volume of information on each job portal, as well as each website’s quality and traffic ranking to select the most important and reliable job portals for an analysis of the labour demand in Colombia. Subsequently, Chapter 5 describes the methodology (web scraping) and different challenges to automatically and rapidly collect a massive number of online job vacancies. The chapter also explains the methods that can be used to homogenise variables such as education and experience and to consolidate information from job portals into a single database.

Next, Chapter 6 illustrates the methods and challenges involved in standardising two of the most relevant variables for the economic analysis of the labour market: skills and occupations. Furthermore, this chapter examines the issues of duplication and missing value, which are some of the main concerns when analysing information from job portals. First, the chapter develops a method to automatically identify skill patterns in job vacancy descriptions based on international skill descriptors and text mining. Then, it proposes
and applies a novel mixed-method approach (software classifiers and machine learning algorithms) to properly classify job titles into occupations. Third, as an employer might advertise the same job many times on the same job portal or on different job portals, the chapter identifies and minimises the issue of duplication. It also explains how missing values were imputed for the “educational requirement” and “wage offered” variables (which are relevant to test the validity of the vacancy database and to analyse labour demand) by using predictors such as occupation, city, and experience requirements. As a result of the above methods, a Colombian vacancy database is generated in Chapter 6 to be tested and analysed to address skill shortage issues.

Subsequently, a comprehensive descriptive analysis of the Colombian labour demand is conducted in Chapter 7. First, the analysis describes the selected job portals, as well as their geographic distribution, in order to build the mentioned vacancy database. Second, it provides a detailed descriptive analysis of the labour demand for skills in Colombia, such as education, occupational structure, potential new occupations, and skills and experience requirements. This description reveals characteristics of the labour demand that were unknown prior to this study. Third, this chapter examines the most notable labour demand trends by occupation: those with higher demand, those with a higher increase, and occupations for which demand has decreased over time. Finally, it describes the distribution of wages offered by employers and other secondary characteristics of the vacancy database, such as contract types and the duration of vacancies.

Although this descriptive analysis might have considerable implications for policymakers and researchers, these results do not provide enough evidence about the validity or reliability of vacancy data to address skill shortages and their consequences. As is the case with data collected by other methods (e.g. surveys), information collected from job portals have limitations that affect interpretation (Chapter 4). Consequently, there is a critical need to assess the validity of the vacancy database to be sure of what it can tell us about labour demand. Thus, Chapter 8 performs extensive internal and external validity tests on the vacancy database (Henson 2001; Rasmussen 2008). First, it evaluates internal validity (consistency of the variables within the same database) via cross-tabulations and wage distribution analysis. Second, it tests external validity (representativeness) of the online vacancy information. This examination
requires a comparison of the vacancy database results against other sources of information (e.g. household surveys). To do so, this book re-categorises occupations from Colombian household surveys to create updated occupational classifications that are compatible with occupational categories in the vacancy database.

After completing the described homologation, a “traditional” test is conducted by comparing the occupational structures of supply and demand. However, given the limitation of the “traditional” test, further tests are carried out to investigate the external validity of the database. Specifically, the wage distribution of labour demand and information on supply are examined to perform a relevant comparison of time series between jobs in demand, employed and unemployed individuals in the total workforce, as well as the extent of new hires (replacement demand and employment growth) by major occupational groups. These detailed tests provide information about the advantages and limitations of the vacancy database for a labour demand and skill mismatch analysis.

Once the advantages and limitations of the data are established, Chapter 9 proceeds to develop a system to identify possible skill shortages and address labour supply according to employer requirements in Colombia. First, the chapter provides a detailed description of the Colombian labour market panorama (formally or informally employed, as well as unemployed) at the occupational level. Second, it combines Colombian household survey information and the vacancy database to estimate a Beveridge curve and a set of eight (volume- and price-based) macro-indicators to identify possible skill shortages. This chapter also highlights the importance of controlling for informality when building skill mismatch indicators in a context such as Colombia. Occupations might exist with relatively low unemployment rates, but also with a relatively high informality rate, or vice versa. Accordingly, increases in the number of workers in certain occupations—for instance, those characterised by relatively low unemployment rates—might increase informality rates. Therefore, this document advises policymakers and training providers to be aware of this relevant labour market duality when providing and promoting skills. Furthermore, this chapter shows how detailed information from vacancies (job descriptions) can be used to monitor labour demand trends for skills, as well as to update occupational classifications according to current employer requirements.
Finally, Chapter 10 summarises the relevant conceptual, methodological, and empirical contributions of the book, while opening a debate on the use of novel sources of information (job portals) to fill information and analysis gaps regarding the labour market. Thus, this chapter highlights the implications of the findings for national statistics offices, policymakers, education and training providers, and career advisers. Additionally, it points out the limitations of the study and illustrates new avenues of enquiry for future research.

This comprehensive and detailed methodological and conceptual framework alongside empirical findings presents important evidence about the advantages and limitations of job portals for their use in economic analysis. It provides a basis to develop a consistent skill shortage monitoring system that can be beneficial for different countries when adopted.