

Received July 30, 2019, accepted August 9, 2019, date of publication August 14, 2019, date of current version September 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2935386

Analysis of the Evolution of the Spanish Labour Market Through Unsupervised Learning

JOSÉ MARÍA LUNA-ROMERA¹, FERNANDO NÚÑEZ-HERNÁNDEZ², MARÍA MARTÍNEZ-BALLESTEROS¹, JOSÉ C. RIQUELME¹, AND CARLOS USABIAGA IBÁÑEZ³

¹Division of Computer Science, University of Seville 41012, Seville, Spain

²Department of Industrial Organization, University of Seville 41092, Seville, Spain

³Department of Economics, Pablo de Olavide University 41013, Seville, Spain

Corresponding author: José María Luna-Romera (jmluna@us.es)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness under Grant TIN2014-55894-C2-R and TIN2017-88209-C2-2-R, in part by the Spanish Ministry of Economics, Industry and Competitiveness under Grant ECO2017-86780, and in part by the Andalusian Government Research Group under Grant SEJ-513 PA.

ABSTRACT Unemployment in Spain is one of the biggest concerns of its inhabitants. Its unemployment rate is the second highest in the European Union, and in the second quarter of 2018 there is a 15.2% unemployment rate, some 3.4 million unemployed. Construction is one of the activity sectors that have suffered the most from the economic crisis. In addition, the economic crisis affected in different ways to the labour market in terms of occupation level or location. The aim of this paper is to discover how the labour market is organised taking into account the jobs that workers get during two periods: 2011-2013, which corresponds to the economic crisis period, and 2014-2016, which was a period of economic recovery. The data used are official records of the Spanish administration corresponding to 1.9 and 2.4 million job placements, respectively. The labour market was analysed by applying unsupervised machine learning techniques to obtain a clear and structured information on the employment generation process and the underlying labour mobility. We have applied two clustering methods with two different technologies, and the results indicate that there were some movements in the Spanish labour market which have changed the physiognomy of some of the jobs. The analysis reveals the changes in the labour market: the crisis forces greater geographical mobility and favours the subsequent emergence of new job sources. Nevertheless, there still exist some clusters that remain stable despite the crisis. We may conclude that we have achieved a characterisation of some important groups of workers in Spain. The methodology used, being supported by Big Data techniques, would serve to analyse any alternative job market.

INDEX TERMS Labour market, cluster analysis, labour mobility, big data.

I. INTRODUCTION

The unemployment rate in Spain is the second highest (15%) among the countries of the European Union after Greece (19%). In recent years the unemployment rate has doubled the European Union average, and the rates are even worse if we focus on youth unemployment, which in 2014 reached 57.9% [1]. Currently, the unemployment rate has a tendency to decline, but it is the cause that most worries to the Spanish, followed by corruption and economic problems [2].

Over recent decades, the Spanish economy has been rooted on a traditional production model based on sectors such as

construction and tourism, which, at the end of the last boom period, accounted for more than a quarter of the national production. In 2008, just at the beginning of the recent economic crisis, the construction sector was around 15% of Spanish GDP -and in addition we would have to take into account the important linked activities-, and tourism was around 11% (in general, the Spanish economy is characterised by an important tertiary bias). Thereby, the collapse of Spain's construction sector -jointly with several related activities- after the bursting of the real estate-financial bubble has beaten records in increasing unemployment at a speed never seen before -the unemployment rate rose from less than 10% to more than 25% in just a few years. Several million jobs were destroyed during the recent economic crisis, going from

The associate editor coordinating the review of this article and approving it for publication was Vijay Mago.

20.6 million employed at the first quarter of 2008 to 16.9 million at the first quarter of 2014 -around 1.7 million lost jobs were in the construction sector. In those years the long-term unemployment problem also regained strength. Fortunately, the labour market figures have moderately improved in the most recent years, although with problems in the quality of the jobs generated.

In the labour market, workers looking for jobs and vacant jobs offered by firms are heterogeneous in many aspects: skills, geographical location, gender, age, payment, etc. These heterogeneities lead to the concept of mismatch: “Mismatch is an empirical concept that measures the degree of heterogeneity in the labour market across a number of dimensions, usually restricted to skills, industrial sector, and location” [3, p. 399].

In this paper, employment data in Spain is processed in order to characterise and to identify groups at the Spanish labour market in order to analyse the evolution that has occurred during and after the crisis. For that reason, we have applied unsupervised machine learning techniques which allow us to discover knowledge from data with just its intrinsic information. In this context, there exists the clustering analysis, that is defined in [4] as the process of partitioning a set of data objects into subsets, where each subset is a cluster; objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. Thereby, clustering is useful in that it can lead to the discovery of previously unknown groups within the data. A clustering analysis is proposed in this study in order to account for the role of heterogeneities in the matching process of the Spanish labour market.

Specifically, we have applied two different clustering algorithms: firstly, partitional and well-known k-means algorithm [5]; secondly, hierarchical clustering with average linkage. One of the main difficulties of cluster analysis is finding the optimal number of clusters. In the k-means algorithm is a prerequisite while in the hierarchical proposal can be decided a posteriori. In this work we have applied both internal and external validation indices to decide the number of clusters in the k-means algorithm, while with average linkage we have opted for a choice in two stages: first we have chosen k based on an internal index and a subsequent refinement based on minimising k with maximum representativeness.

Both clustering techniques have been applied to data from the Spanish labour market in two different economic periods: 2011-2013, which corresponds to an economic crisis period in Spain, and 2014-2016, which has been a period of economic recovery.

The application of both clustering methods to those different economic periods gives four results that are analysed and compared among them. The objective is discuss the evolution of the Spanish labour market over these years of significant economic changes. The main contribution of this article from the labour perspective is to apply unsupervised machine learning techniques to obtain a clearer and more structured information on the employment generation process

and the underlying labour mobility. This information tool, based on the recent labour matching flow, should allow the authorities to orientate, geographically and occupationally, the worker’s search.

The methodology applied in this work is based on Big Data implementations that would allow the analysis performed to be extended to any volume of data regardless of the length of time period analysed or the size of the labour market of a country or international organisations.

The rest of the paper is organised as follows: Section II presents the related works from the literature. Section III establishes the applied methodology including the complete process that is carried out. Section IV details the results, including those that are accomplished by k-means and by average linkage. Finally, Section V summarises the conclusions of our study.

II. RELATED WORK

Data mining is one of the most successful fields of statistics and computer science that uses machine learning, artificial intelligence, statistics and database systems to analyse information in order to discover implicit, new, and potentially useful knowledge from data. Machine learning is the area of artificial intelligence that aims at developing systems that learn automatically and relies on finding patterns and relationships within the data, known as training data, to create models, that is, abstract representations of reality [6]. The training data is composed by a set of examples and each example is characterised by a set of features.

Machine learning tasks are mainly classified into supervised and unsupervised learning. In supervised learning, a mathematical model is created from a set of data that contains the input values and needs a ground truth or prior knowledge of what the output values should be. The most common types of supervised learning are classification (limited set of values for the outputs) and regression (continuous outputs) algorithms. On the contrary, the data only contains input values but does not require labelled output values in unsupervised learning. This kind of algorithms aims to infer the underlying structure or distribution in the data. They can identify patterns or relationships between examples or between features depending on whether they are clustering or association rules algorithms, respectively.

Clustering is one of the most used unsupervised machine learning techniques. Clustering groups the data in clusters so that those data that belong to the same cluster share similar features or attributes, and that data is dissimilar to those in other clusters. The similarity of the data is normally given by how close they are in space, taking into account a distance function [4].

There are many clustering methods in the literature, and there are some works that classify them by some criteria [4], [7], [8]. In this paper we are going to focus on partitional and hierarchical clustering methods. The basic idea of clustering based on partitioning is to divide the data into k groups such that the elements which belong to the same group are

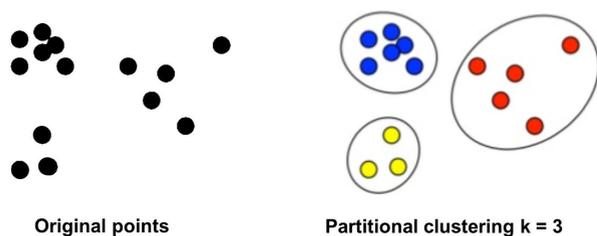


FIGURE 1. Example of clustering based on partitioning methods.

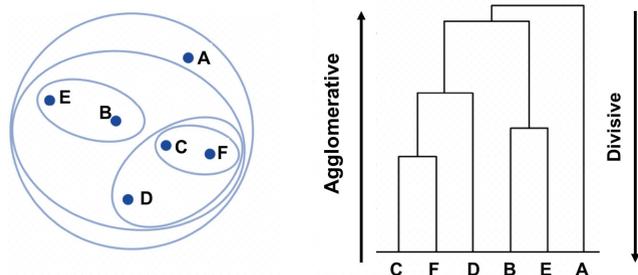


FIGURE 2. Example of clustering based on hierarchical methods.

more similar than the elements from different groups as can be observed in Figure 1. Many partitioning methods form clusters based on distances, so that k clusters are initially assigned, and the object clusters are iteratively changed until a solution where each object is in its nearest cluster is found [9], such as the well-known k -means algorithm [5].

Hierarchical methods create a hierarchical decomposition of the given set of data objects. They can be considered as agglomerative or bottom-up clustering methods if the hierarchy is built assigning each object to its own cluster and then, the most similar clusters are iteratively joined until only a single cluster is left. On the contrary, they can be denoted as divisive or top-down clustering methods if the clusters are created in reverse manner. Thus, all objects are assigned to a single cluster which is recursively split until there is one cluster for each object [10], [11]. The average linkage hierarchical clustering is one of the commonly used hierarchical algorithms where the distance between two clusters is determined by the average distance between each point in one cluster to every point in the other cluster [11].

These methods require a number of clusters into which the data is going to be partitioned. The main problem is that the optimal number of clusters is not known until the clustering is done. This task has been handled in the literature in diverse works [12], [13] establishing the named clustering validity indices (CVI), which are metrics that measure the quality of the clustering. There exists a taxonomy in the literature that distinguishes between two kinds of CVI: internal indices, which measure the quality of the clustering results according to the distance between the clusters, and the compactness of the objects that belong to the same cluster; and external indices, which measure the quality of the clustering solution

through an external indicator of the object distinguishes such as the class.

This paper applies clustering methods to the MCVL information on registered job matches in the Spanish labour market. The nature of our data, with information about jobs and workers having productive matches, links up our work directly with the theoretical concept of the aggregate matching function. This function represents the labour matching process without the need to make explicit the heterogeneities and labour frictions. Instead of representing them specifically according to their origin and their type, heterogeneities and labour frictions are implicitly introduced into an aggregate function that relates the flow of job placements in each period with the levels and inflows of vacancies and job seekers (mainly unemployed seekers). There is an extensive literature (theoretical and empirical) on job search and labour matching processes and, in particular, on the aggregate matching function [3], [14]–[17]. It is important to note that the matching function assumes that workers and jobs are heterogeneous but omits to make those heterogeneities explicit. Without heterogeneities (zero mismatch), the matching function would not exist and jobs and workers would match instantaneously [3], [18]–[20].

Considerable work has been carried out in the literature to open the 'black box' of the matching process and to make explicit the heterogeneities hidden in the matching function. Island models can be found in [21], [22]; urn-ball models in [3]; the taxicab model in [23]; queuing models in [24], [25]; stock-flow models in [26], [27]; and mismatch models in [20]. As a rule, in all these models, workers and jobs are divided into parts (local labour markets, locations, islands, queues, worker-job pairs acceptable or unacceptable to match productively, stock (old)-flow (new) workers and jobs), which are then treated as if each part were homogeneous. Therefore, it is assumed that the heterogeneities of workers and jobs are the reason that the labour market is segmented. Features such as skills, location, age, sex, etc., make certain jobs only suitable for certain workers -there exists evidence of labour market segmentation in the Spanish economy, based primarily on skills and location [28], [29].

The existence of homogeneous groups of workers (and jobs) in a segmented market gives validity to the use of clustering techniques to analyse the matching process in the labour market. Since highly detailed division of the MCVL data in workers or job categories results in a very large number of units, which may be difficult to understand and analyse, we use a clustering methodology, based on a similarity measure, to obtain larger homogeneous groups (clusters) and a better overview of the structure of the labour market [30], [31] which is compatible with the existing theories on labour matching. Cluster analysis enables, as far as possible, subjective or 'a priori' similarity criteria to be avoided -grouping provinces in greater administrative regions, for instance-. Instead, we look for a similarity criterion that is consistent with the search and matching theories applied to labour economics. In this sense, we consider that worker (job)

categories are more similar the more they resemble in the way they match with job (worker) categories; as we shall see, the Manhattan distance is compatible with this idea of similarity in the matching process. It should be highlighted that we have used Manhattan distance based on the works from [32], [33], which used a variant of Manhattan whose values are in the interval $[0, 1]$. Manhattan distance between two worker categories W_i and W_j is defined as:

$$d(W_i, W_j) = \sum_{z=1}^n |W_{iz} - W_{jz}| \quad (1)$$

where n is the total of job categories.

Our study follows the research line of [32], [33] consisting in applying cluster analysis to labour matching data. Other studies have introduced matching data in the analysis of labour clusters. For example, [34] analyses the labour mobility between clusters in Stockholm taking as reference the information and communications technology (ICT) cluster. For these authors, a labour cluster is not simply a large number of firms that belong to the same industrial sector, but a set of complementary and interlinked firms and institutions that have developed a shared consciousness and identity as an industrial cluster and system. In [35] a computer programme that identifies localised mobility clusters in Sweden is developed, the clusters are based on the flows of job movers between workplaces. According to these authors, traditional pecuniary externalities have to be combined with technological and knowledge externalities, coming from the exchange of labour between firms, in order to implement a complete cluster analysis. In this line, the study from [36] used a large Portuguese employer-employee panel-data set to study Marshall's hypothesis that industrial agglomeration improves the quality of firm-worker matching. For these authors, the formation of industrial clusters produces external scale economies, since it increases three intangibles: the potential for more extensive interaction between suppliers and buyers, the firms' ability to capture industry-specific knowledge spillovers resulting from the close proximity of similar firms, and the number of available labour skills and the quality of firm-worker matching. Other articles have analysed labour clusters but without using matching data. For instance, in [37] is studied, for the UK, whether or not different empirical techniques produce identical or similar results in classifying labour markets into homogeneous entities; obtaining some evidence of segmentation in the labour market. The study in [38] worked with a micro-database on workers, for the region of Aragón in Spain, which provides information, among other variables, about where the worker lives and where the worker works. The objective of these authors is to identify local labour markets (clusters) in which a large proportion of the workers both live and work. Meanwhile, following a macroeconomic approach, these works [39]–[41] apply cluster analysis to the Spanish, the European and the German labour market respectively, all of them from a regional perspective. The first authors show that high and

low unemployment Spanish regions have similar responses to regional employment shocks in the short-run, while in the long-run the former are more reactive in terms of spatial mobility. The second paper assesses the impact of the crisis on the Eurozone labour markets integration by conducting a hierarchical cluster analysis. They observe that the last crisis has led to a polarisation of the Eurozone labour markets. Finally, the last study designs a classification approach based on a combination of regression and cluster analysis in order to identify idiosyncratic labour clusters to the Federal Employment Agency. In their two-step methodology, the greater the influence of an exogenous variable on the response variable in the regression analysis, the higher is the weight given to this variable in the cluster analysis. Within all this literature, our work can be inserted into the group of studies that, using labour matching data, generates labour clusters which can be useful for policy-making design and for the management of public employment agencies.

III. PROPOSAL

A. DATASETS

The data used for this purpose comes from the Continuous Working Life Sample (MCVL) [42], a large database containing micro-data on job matches which is provided by the Spanish Ministry of Employment, Migration and Social Security. The MCVL offers information from three Spanish public bodies: labour information from the Social Security system, administrative and personal information from the Continuous Municipal Register of Inhabitants and tax information from the National Tax Agency. The sample is published once a year and the population of reference is composed of individuals who have been paying contributions (such as registered workers or recipients of unemployment benefits) or receiving a contributory pension from Social Security at some date in the year of reference, regardless of how long they have been in that situation. The sample (in each year) comprises 4% of the people belonging to the reference population and is representative of the population registered at the Social Security system in the year of reference. The size of the sample exceeds one million people each year.

In this work, we use the MCVL information to know the characteristics of the workers and the jobs in the job placements that are registered in the Social Security system within the calendar year. The starting point in the processing of the MCVL data is to divide the workers and the jobs involved in the job matches into highly detailed groups according to their characteristics; groups which we call worker categories and job categories, respectively. Ideally, the detailed segmentation should allow us to consider the categories obtained as homogeneous or almost homogeneous, and the large size of the database should enable data (job matches) in each category to be sufficiently numerous as to be statistically representative. Therefore, our unit of analysis (which will be subject to clustering) is not going to be the individual worker

(or the individual vacancy) but its category of belonging. When a job placement occurs, a match is generated between the worker's category and the job's category, a match that may imply a certain degree of occupational or geographical mobility. The availability of appropriate information on geographical and occupational labour mobility is an important requirement for the effectiveness of the labour matching process, and a prominent part of the active labour market policies (ALMPs).

After generating the categories of workers and jobs, the dataset is cross-classified in a contingency table where the rows represent worker categories (WC) and the columns represent job categories (JC). The cells of the contingency table are the frequencies (job matches) between the different categories of workers and jobs; i.e., the cell n_{ij} contains the number of job placements between the worker category w_i and the job category j_j .

As mentioned above, we have applied clustering techniques to two different periods, having each period its own dataset: 2011-2013, which corresponds to a period of economic crisis in Spain; and 2014-2016, which are years of economic recovery. The dataset of the period 2011-2013 contains 5,800 worker categories and 5,198 job categories with a total of 1,967,523 job placements. And the dataset of the period 2014-2016 is composed of 5,722 worker categories and 5,166 job categories with a total of 2,459,686 job placements.

B. METHODOLOGY

The clustering analysis is carried out using two different clustering algorithms and two different technologies: k-means from Spark ML [43], and the average linkage algorithm included in Stata [44]. We have selected these two algorithms because, on the one hand, k-means is one of the most widely used partitioning algorithms, and the Spark version is implemented in a distributed manner and can be executed in a computer cluster. On the other hand, the average linkage is a hierarchical clustering method that has already been widely used in the literature [10], [11], [45], [46]. Therefore, they are widely contrasted clustering techniques and extensively used in many research fields.

These two algorithms have been executed taking the two datasets described in subsection III-A, so we have obtained two clustering results for each dataset. In order to analyse these clustering results, we have followed the methodology from [47]. The first step in a clustering process is to select the optimal number of clusters of each dataset. In the case of k-means from Spark, we have used two kinds of clustering validity indices (CVI), internal and external. We have applied the internal indices BD-Silhouette, BD-Dunn, Davies-Bouldin, and WSSSE included in [13]. In general terms, this kind of CVIs measures how the points are distributed through the clusters taking into account the compactness between the points and the separation between the clusters.

Let Ω be the space of the objects with a given distance d . Then $\{A_k\}_{k=1..N}$ is a set of clusters so that $\bigcup_k A_k = \Omega$, and $A_i \cap A_j = \emptyset \quad \forall i \neq j$.

C_k is the centroid of A_k , and C_0 the centroid of Ω .

Let x_i be an element of A_k , $x_i \in A_k$, and let r_k be the distance from x_i to its own cluster A_k . Then, we can define the following CVIs:

- **BD-Silhouette** (BDS) (Eq 2): This index has been defined, for each possible partition, as the ratio between the difference of the *inter-cluster* and the *intra-cluster* distance, and the maximum of them.

$$BDS = \frac{\text{inter-cluster} - \text{intra-cluster}}{\max\{\text{inter-cluster}, \text{intra-cluster}\}} \quad (2)$$

where *inter-cluster* (Eq 3) is the average of distances between each cluster centroid and the global centroid C_0 :

$$\text{inter-cluster} = \frac{1}{N} \sum_{k=1}^N d(C_k, C_0) \quad (3)$$

and *intra-cluster* (Eq 4) distance is defined as the average of the distances of each point to the centroid of the cluster to which it belongs (Eq 5):

$$\text{intra-cluster} = \frac{1}{|N|} \sum_{x_i \in A_k} r_k \quad (4)$$

where

$$r_k = \frac{1}{|A_k|} \sum_{x_i \in A_k} d(x_i, C_k) \quad (5)$$

BD-Silhouette indicates an optimal value for the number of clusters on the first maximum, which maximises the coherence of the cluster with the lowest possible k .

- **BD-Dunn** (BDD): this index is given, for each possible partition, by the ratio between the minimum of the distances from the centroids to the global centre and the maximum of the distances from each point in the set to its centroid.

$$BDD = \frac{\min_{k=1..N} \{d(C_k, C_0)\}}{\max_{k=1..N} \max_{x_i \in A_k} \{d(x_i, C_k)\}} \quad (6)$$

BD-Dunn points out the number of clusters by the first maximum of the values.

- **Davies-Bouldin**(DB) [48]: In this index, we choose the first minimum of the Davies-Bouldin value chart to create a better model. The index is defined as follows:

$$DB = \frac{1}{N} \sum_i \sum_j \max_{i \neq j} \frac{r_i + r_j}{d(C_i, C_j)} \quad (7)$$

where r_i and r_j are represented in Eq.5, and $d(C_i, C_j)$ is the distance between the centroids C_i and C_j .

- **Within Set Sum of Square Errors** (WSSSE) [43]: This index from Spark ML measures the cohesiveness of the clusters and calculates the sum of the distances from

each point to the centroid of its cluster. The optimal k is generally given by a global minimum or by the result after applying the elbow method to the WSSSE graph.

$$WSSSE = \sum_{x_i \in A_k} d(x_i, C_k)^2 \quad (8)$$

In addition, we have applied the external validity Chi-index to the k-means cluster. This kind of index measures how the points have been distributed by the clusters according to a given class variable. As for the average linkage clustering method, given its hierarchical nature, we have followed an internal validation method to select the optimal number of clusters.

Secondly, we have analysed the clustering results, taking into account the number of elements of each cluster and applying a descriptive statistical analysis. The third step is to evaluate the clustering results based on the features of the points of the datasets in both periods. In our case, we have considered the following features of the worker: region of residence (autonomous community and province), occupation group and sector of activity. Lastly, we have made a comparison between the clustering results for the k-means and the average linkage methods in the two periods.

IV. RESULTS

We have applied k-means from Apache Spark ML [43], and average linkage from [44]. This section includes the results obtained by following the methodology described above. This section is divided as follows: Subsection IV-A includes the clustering analysis using the k-means technique and shows the results for the sub-periods 2011-2013 and 2014-2016. Subsection IV-B follows the same structure of the previous subsection but the results are those of the average linkage method. Each of these subsections includes the selection of the optimal number of clusters, the description of the clustering results, and a comparison between the results of both sub-periods. Finally, Subsection IV-C carries out a comparison between the results of the k-means clustering and those coming from the average linkage clustering.

A. K-MEANS

Figure 3 shows the results of the internal CVIs of the k-means cluster for the sub-period 2011-2013. Each index is interpreted differently: Silhouette follows the “elbow method”, which establishes the optimal number of clusters when the curve of the index begins to stabilise. In this case, Silhouette does not stabilise at any point until $k = 50$. The Dunn index points out the optimal number of clusters with local minimums; in this case, we can observe some local minimum along the curve, but we may not conclude that they are proper solutions because they are not decisive enough. Davies-Bouldin index points out the optimal number of clusters with local maximum, and as happened with the Dunn value, there are some local maximum but they do not look like suitable solutions because there are not determinant numbers. Finally, the WSSSE function points out the optimal solution

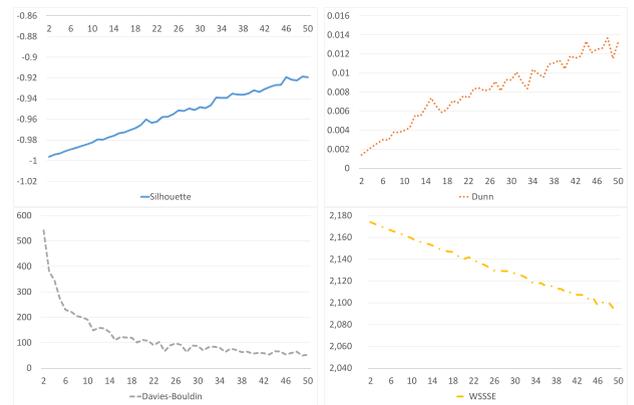


FIGURE 3. Internal CVI of k-means in the period 2011-13. X-axis represents the number of clusters and Y-axis the value of the index.

as Silhouette, but the other way around, and we cannot find any stabilisation until $k = 50$. As can be observed, none of the indices concludes with an optimal number of clusters -we have found a similar situation for the next sub-period (2014-2016), so we have omitted the inclusion of the corresponding figure- so external CVIs need to be applied in order to find a proper clustering solution.

Figure 4 shows the results of the Chi Index [49] for the sub-period 2011-2013. Chi Index is defined as an external CVI which measures the quality of a clustering by means of the distribution of the instances through the classes, and the classes through the clusters. Chi Index measures the coherence between a class variable and a cluster through a contingency matrix. This matrix denotes the number of elements (job matches in our case) of each cluster (rows) in each value or category of the class variable (columns) in such a way that each cell ij of the matrix shows the total of matches of the cluster i in the category j of the class variable. Chi Index measures the coherence of this matrix dividing it into two components: the first one is a contingency matrix with relative values with respect to the marginal distribution of the clusters (represented by the blue line); the second one is the contingency matrix with relative values taking the marginal distribution of the class variable as reference (represented by the orange line). In this way, Chi Index is represented by two curves, which are the ones shown in the corresponding graphs of Figure 4. Chi Index was calculated assuming as classes: the region (Spanish Autonomous Communities), the province, the occupation group and the activity sector of the worker. Chi Index points out the optimal number of clusters (k) in the intersection between the curves.

We have decided not to show the graphs of the external index for the sub-period 2014-2016 because of space limitations; however, the results of this sub-period are included in Table 1. Table 1 represents the results of the Chi Index by each class in each sub-period. We have selected as the optimal number of clusters the one given by the region (rejecting province, occupation and sector classes) because it includes the province by definition, so that the province is directly

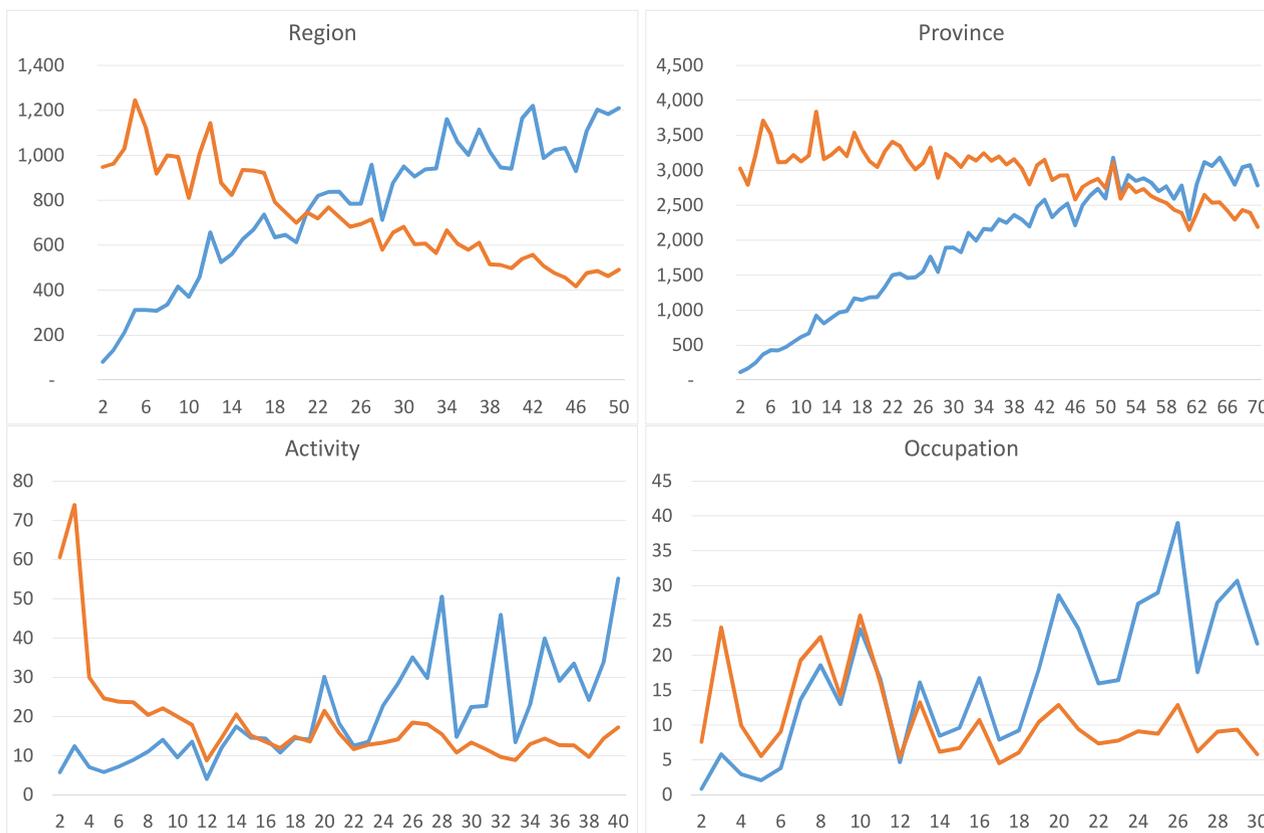


FIGURE 4. External clustering validity indices for k-means in the period 2011-13. X-axis represents the number of clusters. Y-axis shows the value of the index.

TABLE 1. Results of the external validity clustering indices for the periods 2011-13 and 2014-16. In bold, the chosen result.

Class	2011-2013	2014-2016
Region	21	22
Province	51	53
Activity	16	13
Occupation	11	10

located inside the region. In addition, the number of clusters given by the province (51 and 53 in each sub-period) is too large for having an easy to read and handle cluster solution. On the other hand, the activity and the occupation obtained lower numbers of clusters than the region, so, we may assume that these solutions are also included in the optimal number of clusters given by the region. Hence, we have considered $k = 21$ for the sub-period 2011-2013, and $k = 22$ for the next sub-period, 2014-2016.

Table 2 shows the number of worker categories and job placements for the clusters $k = 21$ means in the sub-period 2011-2013, and $k = 22$ means in the sub-period 2014-2016. It is worth mentioning that the clusters with the same identification number in both sub-periods are not the same, the number is just used to name them. In addition, it must be observed that the sub-period 2014-2016 has got one cluster more than the previous sub-period.

Focusing on the sub-period 2011-2013, the 5,800 WCs have been homogeneously distributed across all the clusters.

TABLE 2. Clustering results for k-means with $k = 21$ in the period 2011-13 and with $k = 22$ in the period 2014-16. Minimum and maximum of each column are highlighted in bold.

#	2011-2013		2014-2016	
	WCs	Placements	WCs	Placements
0	331	74,545	229	56,486
1	446	174,900	350	213,003
2	176	73,095	225	2,226
3	391	237,447	173	33,686
4	414	197,624	142	57,341
5	132	51,316	195	61,578
6	430	43,664	340	67,956
7	256	39,617	205	74,264
8	306	99,857	308	118,065
9	342	56,912	114	8,624
10	207	80,621	247	111,389
11	267	48,643	289	152,245
12	341	159,071	196	85,523
13	124	30,808	343	199,403
14	122	31,493	183	28,121
15	213	71,198	244	65,130
16	141	8,024	379	147,638
17	250	132,465	243	173,358
18	385	268,371	400	350,751
19	298	63,395	602	388,434
20	228	24,457	133	30,246
21	-	-	182	34,219
Total	5,800	1,967,523	5,722	2,459,686
Avg	276.19	93,692	260.09	111,803.91
Min	122	8,024	114	2,226
Max	446	268,371	602	388,434

Clusters have an average size of 276 WCs. Cluster 14 is the smallest one with 122 WCs (2% of the total), and cluster 1 has

TABLE 3. Summary of the clustering features of k-means with $k = 21$ in the period 2011-13.

#	Size	Location	Activity	Occupation
0	M	Centre	Serv3 / Serv2	TechEngin / UnivDegr / C&WHeads
1	L	North-East	PublicAd / Serv1 / Constr / Manuf	Low-skilled
2	S	South	Agric	Low-skilled
3	L	South / Ceuta	Constr / PublicAd / Serv3	Low-skilled
4	L	Centre	PublicAd / Educ / Health	Low-skilled
5	S	Balearic I.	Supplies / PublicAd / Serv1	Low-skilled
6	L	North / South / Centre	Manuf / Supplies	TechEngin / Low-skilled
7	M	Centre	Agric / Health / Serv1	Low-skilled
8	M	Northeast / Northwest	PublicAd / Educ / Serv3	Low-skilled
9	M	Centre	PublicAd / Constr / Educ / Serv3	Low-skilled
10	S	Canary I.	PublicAd / Serv1 / Educ / Serv3	Low-skilled
11	M	Centre / East	Serv1	Low-skilled
12	M	South / East	Manuf	Low-skilled
13	S	Centre	Constr	Low-skilled
14	S	North	Serv2	C&WHeads
15	S	North	Educ / Manuf / Serv3 / Health	Low-skilled
16	S	Centre / South	Serv1 / Manuf / Serv2	TechEngin / C&WHeads
17	M	Centre / South	Agric	Low-skilled
18	L	Northeast	Manuf	C&WHeads / Low-skilled
19	M	Northwest / Centre	PublicAd / Serv1 / Constr	Low-skilled
20	S	Centre	PublicAd / Educ	TechEngin / UnivDegr

TABLE 4. Summary of the clustering features of k-means with $k = 22$ in the period 2014-16.

#	Size	Location	Activity	Occupation
0	M	North / Centre	Agric	Low-skilled
1	L	East / Canary I.	Constr	Low-skilled
2	S	Northwest / Centre	Supplies / ExtratOrg	Low-skilled / UnivDegr / C&WHeads
3	S	Centre	Serv1 / PublicAd	Low-skilled
4	S	South	Educ / Health	TechEngin / UnivDegr
5	S	North	Educ / Health / Manuf	Low-skilled
6	L	North / Centre	Serv3	TechEngin / UnivDegr
7	S	Centre	Agric	Low-skilled
8	M	North	Manuf / Serv2	Low-skilled
9	S	Northeast	Serv2	TechEngin / UnivDegr / C&WHeads
10	M	Centre	Constr	Low-skilled
11	M	South	Agric / Educ / Serv1	Low-skilled
12	S	Centre / Canary I.	Serv1 / Constr	Low-skilled
13	M	Centre	Serv1 / Manuf / Serv3 / Serv2	TechEngin / C&WHeads / Low-skilled
14	S	Northeast	PublicAd / Serv1	Low-skilled
15	M	Centre	Educ	Low-skilled
16	L	Northwest	PublicAd	Low-skilled
17	M	East	Construc / Manuf	Low-skilled
18	L	Northwest	PublicAd / Serv1 / Serv3	Low-skilled
19	L	South / Ceuta / Balearic I.	PublicAd / Constr / Serv1	Low-skilled
20	S	Centre	Health	TechEngin / C&WHeads
21	S	Centre	Health / Serv1	Low-skilled

got the highest number of elements, with 446 WCs (11%). In general terms, the job placements are in line with the size of the cluster, with the largest group being the one with the largest number of job placements. On the other hand, the result for $k = 22$ in the sub-period 2014-2016 does not differ very much from the previous scenario. As can be observed, the clusters of this sub-period are composed of 260 WCs on average, with a range between 114 WCs (cluster 9) and 602 WCs (cluster 19), and between 2,226 matches (cluster 2) and 388,434 matches (cluster 19).

A summary of the cluster structure by region, province, activity sector and occupation group can be found in the Tables 3 and 4, one for each sub-period. These tables are built by considering for each cluster only those categories of the variables with the highest percentages in terms of job matches. Specifically, they show the id number of the cluster;

the size of the cluster in terms of job matches, which was set in intervals of the equal width, starting with the size of the smallest cluster (114), so that, 'S' is set for small clusters in the range [114, 228], medium (M) within the range (229, 343], and large (L), for those clusters larger than 343 elements; the main locations of the cluster in cardinal points form; the ids of the main sectors of activity, whose respective assignment can be found in Table 12; and the main occupation groups of the clusters which have been grouped in the following categories: Managers and workers with university degree (UnivDegr), Technical engineers and qualified assistants (TechEngin), Clerical and workshop heads (C&WHeads), and the rest of occupations, which have been categorised as Low-skilled.

Table 3 shows the clustering features for the sub-period 2011-2013. As can be seen, there exist five large clusters (1, 3, 4, 6 and 18) geographically distributed in different

spatial areas of the country. It is interesting to note that these clusters are mainly composed of low-skilled workers, with the exception of clusters 6 and 18 which add Technical Engineers and C&WHeads respectively. In addition, there is no predominant sector of activity, although the most common sector is the manufacturing industry, which is present in 3 of these 5 clusters. It should be highlighted that clusters 5 and 10 are mainly based on workers from the Canary Islands and the Balearic Islands respectively. Besides that, the clusters 0, 6, 16 and 20 are the only ones with Technical Engineers, and just the clusters 0 and 20, which are located in the Centre of Spain, are in addition composed of University degrees. It is noteworthy that agricultural workers are mainly located in the Centre and the South of the country, as the clusters 2, 7, and 17 show. It is also interesting to point out that there are four clusters (0, 14, 16, and 18) whose principal occupation group is C&WHeads; they mainly share the Financial & Business Services activity and do not have a predominant location.

Table 4 summarises the features of the clustering for the sub-period 2014-2016. In general, the clusters are from just one location, and when there is more than one location, they show geographical proximity. The largest clusters are mainly composed of worker categories from the North but the cluster 19 that is composed of Southern (including Ceuta) and Balearic workers. All these clusters have a non-qualified occupation group as predominant, except the cluster 6 that is just of higher education. In this period, there are more small clusters than medium ones, and the cluster size is related neither to the occupation group nor to the sector of activity. In this sub-period, there are also three clusters (4, 6, and 9) whose occupation group is mainly composed of workers with university studies; the main location of these clusters is the North of the country and they do not share any specific economic activity.

1) 2014-16 VS 2011-13

This section carries out the comparison between the clusters of the sub-periods 2011-2013 and 2014-2016. Table 5 shows the clusters from the sub-period 2014-2016 and their correspondence with the clusters of the previous sub-period in terms of the worker categories that they have in common. The colour of the cells indicates the level of relationship between the clusters, the darker the green, the stronger the relationship-the greater is the number of the worker categories that they have in common.

It should be highlighted that there are six clusters (3, 4, 10, 15, 17, and 20) which have correspondence only with one of the clusters of the previous sub-period, although this correspondence is not 100% or one-to-one, since the corresponding clusters of the first sub-period (2011-2013) with which the six clusters match also appear related to other clusters of the second sub-period analysed (2014-2016); in other words, some clusters of sub-period 2011-2013 have been separated into different clusters of the sub-period 2014-2016. For instance, the cluster 17 of the second sub-period (2014-2016), which is mainly composed of

TABLE 5. Correspondence between the k-means clusters of the periods 2014-16 and 2011-13.

2014-16	2011-13		
0	6	19	
1	1	10	11
2	6		
3	11		
4	2		
5	1	15	
6	7	20	1
7	2	7	
8	15	14	1
9	0	18	9
10	4		
11	17	12	6
12	10	0	
13	4	0	
14	18	8	
15	9		
16	8	19	
17	12		
18	18	8	
19	3	5	
20	20		
21	13	7	

low-skilled individuals working at construction and industrial manufacturing in the East side of the country, belongs to a larger cluster (the number 12) of the first sub-period (2011-2013) which also keeps some correspondence with the cluster 11 (2014-2016); cluster 12 (2011-2013) is a cluster of low-skilled workers from the industrial manufacturing sector. We find a similar result with cluster 15, which belongs to cluster 9 of the sub-period 2011-2013; in both clusters we find low-skilled workers in the central area and working in the educational sector.

On the other hand, the clusters 1, 6, 8, and 9 of sub-period 2014-2016 are linked with multiple clusters of the first sub-period (2011-2013). This result may indicate that larger local labour markets have emerged in this second sub-period. For example, the cluster 8 from sub-period 2014-2016, which is mainly formed by low-skilled workers from the Northern area in the industrial manufacturing, and financial and business services, is composed of worker categories of the clusters 1, 14 and 15 of sub-period 2011-2013, which are located in the North, with low-skilled workers in most cases and with a range of different economic activities. In the case of the cluster 9, which is composed of high education workers from the Northeast in the activity of financial and business services, is formed of clusters 0, 18 and 9 of the first sub-period; these clusters mainly have workers with higher education, are also dedicated to the financial and business services and share some geographical locations.

The rest of the clusters (0, 5, 7, 12, 13, 14, 16, 18, 19 and 21) are related to two clusters of the first sub-period, although only with one of them maintain a strong relationship. For instance, cluster 5, which is composed of low-skilled workers from the Northern area and is dedicated to education, health and industrial manufacturing, is formed by clusters 1 and 15 of the sub-period 2011-13, which are clusters from the

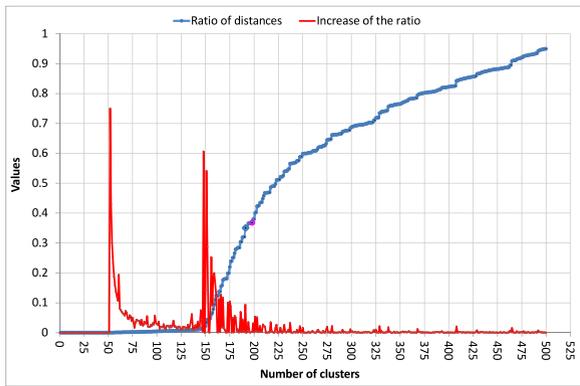


FIGURE 5. Representation of the selection of the optimal number of clusters in the period 2011-13. The blue line represents the ratio between inter-cluster and the intra-cluster distances for each possible number of clusters, and the red line represents the increase of that ratio in percent. The chosen optimal number of clusters was 191 of which we have studied 23.

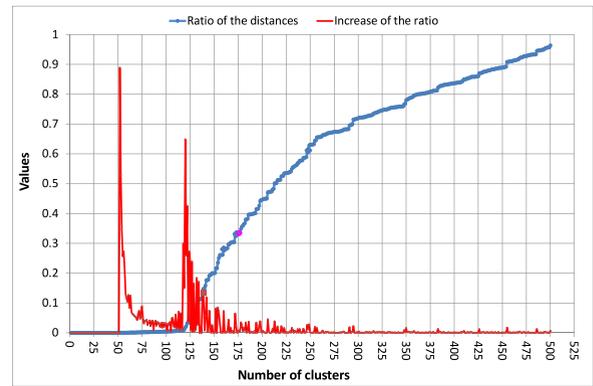


FIGURE 6. Representation of the selection of the optimal number of clusters in the period 2014-16. The blue line represents the ratio between inter-cluster and the intra-cluster distances for each possible number of clusters, and the red line represents the increase of that ratio in percent. The chosen optimal number of clusters was 176 of which we have studied 25.

North, and share similar economic activities and occupation groups. We find a similar situation with cluster 16, which is from the Northwest, dedicated to public administration, and it is composed of clusters 8 and 19 of the sub-period 2011-13, which are also from the Northwest and belong to the public administration sector. Another case is the one of cluster 19 from 2014-16 (large size), which is mainly composed of cluster 3 and, to a lesser extent, of cluster 5. These clusters are located in the South, Ceuta and Balearic Islands, and belong to the public administration and construction sectors.

We can conclude that due to the change in the cycle of the economy, there have been some movements in the Spanish labour market which have changed the physiognomy of some of the ‘job creation’ clusters. However, there still exist some clusters that remain stable despite the economic crisis (showing some degree of inertia).

B. AVERAGE LINKAGE

This section follows a similar structure than the previous one. Firstly, it contains the study of the optimal number of clusters for both sub-periods, and then, the optimal clustering result is described.

Figure 5 relates the inter-cluster and the intra-cluster distances for each possible number of clusters. The blue line in the figure represents the ratio between both distances, and the red line represents the increase of that ratio in percent. In this case, we have chosen $k = 191$ as the optimal number of clusters because a proper solution is given by a highest ratio between the inter-cluster and the intra-cluster distances until its increment stop raising, so that the red line tends to zero. After the selection of $k = 191$, we have only taken the 23 clusters that have 1,500 or more job placements. In this way, we keep 99% of the job placements, and just skip those clusters which contain few elements. We must bear in mind that choosing the 23 largest clusters for $k = 191$ is not the same as initially estimating $k = 23$.

In the same way, Figure 6 shows the results for the period 2014-16, where we have taken $k = 176$ as the optimal number of clusters; of those clusters, we have analysed the 25 largest-those with more than 1,500 job placements-.

Table 6 shows the results for the average linkage with $k = 23$ and $k = 25$ in the sub-periods 2011-13 and 2014-16, respectively. The data analysed with the average linkage method for the first sub-period is composed of a total of 5,317 worker categories that give rise to 1,941,816 job matches. The 23 clusters have got 231 WCs and more than 84,000 job placements on average. The cluster 41 is the one with the fewest number of WCs, just 42, and cluster 11 is the one with the fewest number of matches (4,940). As can be observed, the clusters with more WCs and job placements do not match either: cluster 1, which has the largest number of WCs, contains 801 WCs with 289,446 job placements, while cluster 12, the one with the largest number of matches, is composed of 531 WCs with 304,996 job placements.

On the other hand, 5,486 worker categories are analysed during the sub-period 2014-16. In this case, the clusters have 219 WCs and 98,257 job placements on average. Cluster 7 is the one with more WCs and placements, 878 and 388,935 respectively. Moreover, the clusters with the lowest number of WCs and placements do not match: cluster 4 contains only 3 WCs and 1,815 matches, and cluster 24 has 43 WCs and just 1,640 matches.

Table 7 summarises the features of the clustering result for the sub-period 2011-13 with $k = 23$. In this case, the size of the clusters is evenly divided. There is just one large cluster, 7 medium clusters and 15 small clusters. It should be highlighted that there are 11 clusters which are composed of worker categories from the centre of Spain. There are just 4 clusters (2, 3, 8, and 41) with university studies as principal occupation group, of which three of them are located in the Centre and the South of the country, and their main activities are health, manufacturing and some services. Likewise, there are two clusters (11 and 20) whose main occupation

TABLE 6. Clustering result for average linkage with $k = 23$ and $k = 25$ in the periods 2011-13 and 2014-16, respectively. Minimum and maximum of each column are highlighted in bold.

2011-2013			2014-2016		
#	WCs	Placements	#	WCs	Placements
1	801	289,446	1	296	120,188
2	246	16,683	2	87	28,378
3	75	30,405	3	221	60,089
4	186	47,340	4	3	1,815
6	189	55,131	7	878	388,935
8	86	10,855	9	192	70,966
9	69	5,543	10	23	2,328
10	80	20,123	11	42	18,352
11	67	4,940	13	242	42,833
12	531	304,996	14	133	18,880
14	307	119,781	15	115	41,140
16	198	90,041	16	110	17,666
19	489	298,560	17	381	201,040
20	108	49,342	18	228	160,443
21	289	53,035	19	214	103,842
22	210	123,536	21	312	233,901
23	222	110,822	22	521	282,188
24	164	26,685	24	43	1,640
25	262	94,269	25	492	375,928
27	93	13,007	26	122	62,932
29	205	46,134	28	270	65,786
34	398	112,161	32	385	126,576
41	42	18,981	34	64	3,407
-			38	48	23,699
-			44	64	3,482
Total	5,317	1,941,816	Total	5,486	2,456,434
Avg	231.17	84,426.78	Avg	219.44	98,257.36
Min	42	4,940	Min	3	1,640
Max	801	304,996	Max	878	388,935

TABLE 7. Summary of the clustering features of average linkage with $k = 23$ in the period 2011-13.

#	Size	Location	Activity	Occupation
1	L	Centre	Serv3 / Educ	Low-skilled
2	M	Centre / South	Serv1 / Manuf / Serv2	UnivDegr / TechEngin
3	S	Centre / South	Health	TechEngin
4	S	North / Centre	Health / Constr / Serv3	Low-skilled
6	S	West	Agric / Educ / PublicAd	Low-skilled
8	S	Centre	Manuf / Serv1 / Serv3	TechEngin
9	S	Centre	Serv1 / Health / Constr	Low-skilled
10	S	North / Centre	Manuf / Constr	Low-skilled
11	S	Centre	Serv1 / PublicAd	C&WHeads
12	M	South	PublicAd / Agric / Serv3	Low-skilled
14	M	South	Agric / Serv1	Low-skilled
16	S	Canary I.	Serv1 / Serv3	Low-skilled
19	M	Northeast	PublicAd	Low-skilled
20	S	Balearic I.	Serv1 / Serv3	C&WHeads
21	M	Northeast	PublicAd / Serv3	Low-skilled
22	S	East	Manuf / Serv3	Low-skilled
23	S	East	Constr / Manuf	Low-skilled
24	S	Centre	Serv1	Low-skilled
25	M	North	Educ / Serv3 / Health	Low-skilled
27	S	Centre	Health / Serv1 / Educ	Low-skilled
29	S	North	Constr / Manuf	Low-skilled
34	M	Northwest	Constr	Low-skilled
41	S	Centre / South	Health	UnivDegr

group is C&Wheads (the main activity is trade, transport, accommodation and communication), but one is placed in the Centre and the other in the Balearic Islands. The rest of the clusters (17 clusters) have no high education levels among their main occupations: three of them (6, 12, and 14) are based on agriculture (West or South location and medium or small-size); other one (small) is composed of Canary workers in the sector of trade, transport, accommodation, communication, and other services; and two of them (22 and 23) are from the

TABLE 8. Summary of the clustering features of average linkage with $k = 25$ in the period 2014-16.

#	Size	Location	Activity	Occupation
1	M	North	Health / Serv1	Low-skilled
2	S	North	Manuf / Constr	Low-skilled
3	S	North	Agric	Low-skilled
4	S	North	Health / Constr / Serv3	TechEngin / C&WHeads
7	L	Centre	Constr / Educ / Serv3	Low-skilled
9	S	West	Serv1 / PublicAd	Low-skilled
10	S	Centre / South	PublicAd	UnivDegr
11	S	Centre / South	Health	TechEngin / UnivDegr
13	S	Centre	Serv1	Low-skilled
14	S	Centre	Serv1 / PublicAd	Low-skilled
15	S	North	Serv2 / Constr	Low-skilled
16	S	Centre	Health / Manuf / Serv1	Low-skilled
17	M	Centre / East	Manuf / Serv1	Low-skilled
18	S	East	Educ	Low-skilled
19	S	Canary I.	Serv1	Low-skilled
21	M	South	Constr / Agric	Low-skilled
22	L	South	Educ / Agric	Low-skilled
24	S	Ceuta	PublicAd / Serv1	Low-skilled
25	L	Northeast	PublicAd	Low-skilled
26	S	Balearic I.	Serv1 / Educ / Manuf	Low-skilled
28	M	Northeast	PublicAd / Serv1	Low-skilled
32	M	Northwest	Serv1 / PublicAd	Low-skilled
34	S	Centre	Agric	Low-skilled
38	S	All	Health	UnivDegr
44	S	Melilla	Serv1 / PublicAd	Low-skilled

East and share the industrial manufacturing sector as main economic activity, among others.

Next, the features of the clustering for the sub-period 2014-16 are going to be discussed (Table 8). In this case, we find 3 large clusters, 5 medium clusters, and 17 small clusters. The clusters 4, 10, 11 and 38 are the only ones with university studies; in addition, their main sector of activity is health, and they are located all around Spain. The other clusters do not have, in general, high level of studies. In these clustering results, we find several clusters located only in one province, such as the 19 (Canary Islands), the 26 (Balearic Islands), the 24 (Ceuta), and the cluster 44 (Melilla); all of them are mainly focused on the sector of trade, transport, accommodation and communication. The average linkage clustering in this sub-period also includes two clusters (21 and 22) from the Southern zone whose principal economic activity is agriculture, although they also include workers of the sectors of construction and education.

C. K-MEANS VS AL CLUSTERS

This section includes a comparison between the results of the k-means (KM) and the average linkage (AL) methods in both sub-periods. Tables 9 and 10 show the correspondence between the results of the k-means and the average linkage clusters during the periods 2011-13 and 2014-16 respectively. As mentioned above, the colour of the cells indicates the level of relationship between those clusters, the darker the green, the stronger the relationship.

Table 9 shows the comparison for the sub-period 2011-2013. The KM clusters 3, 4, 5, 9, 10, 13, 14, 15, 16, and 18 are directly related with just one AL cluster. This indicates that the clustering results of the KM are similar to those of the average linkage. There are 6 KM clusters that are composed of two AL clusters, but only with one of them, the relationship

TABLE 9. Correspondence between the 21 clusters of k-means and 23 clusters of AL in the period 2011-13.

K-means	Average Linkage		
	1	2	8
0	1	2	8
1	29	22	10
2	6	12	
3	12		
4	1		
5	20		
6	4	12	14
7	1	6	9
8	34	19	
9	21		
10	16		
11	24	22	
12	23	14	
13	1		
14	25		
15	25		
16	2		
17	14	12	
18	19		
19	34	4	
20	27	11	1

TABLE 10. Correspondence between the 22 clusters of k-means and 25 clusters of AL in the period 2014-16.

K-means	Average Linkage		
	15	16	
0	15	16	
1	18	19	
2	32	7	
3	17		
4	21		
5	3	1	
6	13	7	2
7	9		
8	1	3	
9	25	7	3
10	7		
11	22		
12	19	14	
13	7		
14	25	28	
15	28	34	
16	32		
17	17		
18	25		
19	22	21	26
20	13	11	
21	7	14	

TABLE 11. Similarity between the clustering results of the k-means and the average linkage.

Similarity	1 Cluster	2 Clusters	3 Clusters
2011-2013	66%	83%	90%
2014-2016	71%	91%	98%

can be considered strong. Just the KM clusters 0, 6, and 20 have got a weak relationship with the AL cluster.

We find a similar situation in the second sub-period (2014-2016), which is represented in Table 10. The KM clusters 3, 4, 7, 10, 11, 13, 16, 17 and 18 are directly related with just one AL cluster. Likewise, the KM clusters 0, 1, 5, 8, 12, 14, 15, 20, and 21 are composed of two AL clusters; i.e., the worker categories of some AL clusters are joined to

TABLE 12. List of sectors of activity with their assigned code.

Activity sector	Id
Agriculture	Agric
Construction	Constr
Education	Educ
Extraterritorial Organisations	ExtratOrg
Health	Health
Manufacturing	Manuf
Mining	Mining
Public Administration	PublicAd
Trade, Transport, Accommodation & Communication	Serv1
Financial & Business Services	Serv2
Other Services	Serv3
Supplies	Supplies

build a new KM cluster. Finally, there are just 3 KM clusters (2, 9 and 19) that do not have a strong relationship with any specific AL cluster; they match with several AL clusters but at a very low rate.

In order to quantify the similarity between the clustering solutions (KM and AL) of each sub-period, we have calculated the ratio of coincidence between those solutions. For that purpose, we have considered the similarity between a KM cluster and an AL cluster as the ratio of the elements (worker categories) in common in relation to the total of elements of the KM cluster. This comparison can also be done on a scale of one-to-many (one KM cluster and several AL units), so, for each KM cluster, we have progressively taken from 1 to 3 AL clusters (sorted from the highest to the lowest relation with the KM cluster) in order to calculate the corresponding ratios. Table 11 shows the results of our comparison. The different sub-periods are represented by rows, and the number of clusters that we have taken to make the comparison is expressed by columns. In the period 2011-2013, we have obtained a 66% of similarity between the KM clusters and AL clusters taking just the AL unit which has the highest number of common worker categories. Taking 2 AL clusters, we have obtained that the clusters are similar by 83%. Finally, taking 3 AL clusters, we have obtained a similarity rate of 90%. Furthermore, we find a similar picture for the period 2014-2016, but with higher percentages, ranging from 71% of similarity if we take the AL cluster with the highest rate to 98% if we take 3 AL clusters.

V. CONCLUSION

In this study, a labour matching analysis of the Spanish labour market is developed based on the recent labour matching flow. This analysis may allow the authorities to orientate, geographically and occupationally, the worker's search. We have applied an unsupervised machine learning technique, such as the clustering methodology, with the aim to discover how the labour market is organised, taking as unit of analysis the different categories of the workers who get a job. The initial databases have been pre-processed to work with worker and job categories which are related through a contingency table that contains the job placements that occur between them, representing a two-sided matching model. We have applied two different clustering algorithms, with

different technologies. Thereby, with each clustering algorithm, we have applied different methods to discover the optimal number of clusters. Then, we have characterised the clustering results, focusing on the size of the clusters, the geographical location, the activity sector, and the occupation group of the workers. Finally, we have made a comparison between the different periods to see the evolution of the labour market under both clustering methods. Our methodology is versatile and could be adapted to many other labour analyses.

The findings of this study provide evidence of the effects of the recent economic crisis in the Spanish labour market. One could conclude from these results that there have been some transformations in the Spanish labour market, which have changed the physiognomy of some of the “job creation” clusters. However, there still exist some clusters that remain stable despite the economic crisis. These movements have been observed in the results of both clustering methods. In addition, there exists a strong similarity between the k-means and average-linkage results, in such a way that the ratio of similarity was between the 66% and 98% depending on the number of AL clusters that we take into account. These two approaches can support the economic and political decision making in different public administrations, as well as the customisation of the employment policies, improving the ALMPs.

Finally, we have also achieved an interesting characterisation of groups of workers all around Spain. In this sense, our methodology is also useful to capture the structure of the labour market (local labour markets, for instance).

ACKNOWLEDGMENTS

J.M. Luna-Romera holds a FPI scholarship from the Spanish Ministry of Economy and Competitiveness. *José María Luna-Romera and Fernando Núñez-Hernández contributed equally to this work.*

REFERENCES

- [1] European Union. (2018). *Total Unemployment Rate*. Accessed: Oct. 18, 2018. [Online]. Available: <https://ec.europa.eu/eurostat>
- [2] (2018). *Three Principal Problems*. Accessed: Oct. 18, 2018. [Online]. Available: <http://www.cis.es/>
- [3] B. Petrongolo and C. A. Pissarides, “Looking into the black box: A survey of the matching function,” *J. Economic Literature*, vol. 39, no. 2, pp. 390–431, 2001.
- [4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.
- [5] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1967, pp. 281–297.
- [6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.
- [7] D. Xu and Y. Tian, “A comprehensive survey of clustering algorithms,” *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015.
- [8] Nisha and P. J. Kaur, “A survey of clustering techniques and algorithms,” in *Proc. 2nd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, 2015, pp. 304–307.
- [9] J. Han, M. Kamber, J. Pei, J. Han, M. Kamber, and J. Pei, “10—Cluster analysis: Basic concepts and methods,” in *Data Mining*. San Mateo, CA, USA: Morgan Kaufmann, 2012, pp. 443–495.
- [10] A. Triayudi and I. Fitri, “Comparison of parameter-free agglomerative hierarchical clustering methods,” *ICIC Express Lett.*, vol. 12, no. 10, pp. 973–980, 2018.
- [11] B. Moseley and J. Wang, “Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 3094–3103.
- [12] J. Wu, J. Chen, H. Xiong, and M. Xie, “External validation measures for K-means clustering: A data distribution perspective,” *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6050–6061, 2009.
- [13] J. M. Luna-Romera, J. García-Gutiérrez, M. Martínez-Ballesteros, and J. C. R. Santos, “An approach to validity indices for clustering techniques in big data,” *Prog. Artif. Intell.*, vol. 7, no. 2, pp. 81–94, 2018.
- [14] T. J. Devine and N. M. Kiefer, *Empirical Labor Economics: The Search Approach*. New York, NY, USA: Oxford Univ. Press, 1991.
- [15] D. Mortensen and C. Pissarides, “New developments in models of search in the labor market,” in *Handbook Labor Economics*, vol. 3, O. Ashenfelter and D. Card, Eds., 1st ed. Amsterdam, The Netherlands: Elsevier, 1999, pp. 2567–2627.
- [16] R. Rogerson, R. Shimer, and R. Wright, “Search-theoretic models of the labor market: A survey,” *J. Econ. Literature*, vol. 43, no. 4, pp. 959–988, 2005.
- [17] E. Yashiv, “U.S. labor market dynamics revisited,” *Scandin. J. Econ.*, vol. 109, no. 4, pp. 779–806, 2007.
- [18] C. A. Pissarides, *Equilibrium Unemployment Theory*. Cambridge, MA, USA: MIT Press, 2000.
- [19] C. Pissarides, “Economic dynamics interviews Christopher Pissarides on the matching function,” in *Economic Dynamics Newsletter*, vol. 10, no. 1. Amsterdam, The Netherlands: Elsevier, 2008.
- [20] R. Shimer, “Mismatch,” *Amer. Econ. Rev.*, vol. 97, no. 4, pp. 1074–1101, 2007.
- [21] R. Lucas and E. Prescott, “Equilibrium search and unemployment,” *J. Econ. Theory*, vol. 7, no. 2, pp. 188–209, 1974.
- [22] D. T. Mortensen, “Island matching,” *J. Econ. Theory*, vol. 144, no. 6, pp. 2336–2353, 2009.
- [23] R. Lagos, “An alternative approach to search frictions,” *J. Political Economy*, vol. 108, no. 5, pp. 851–873, 2000.
- [24] P. A. Gautier, “Non-sequential search, screening externalities and publications and the public good role of recruitment offices,” *Econ. Model.*, vol. 19, no. 2, pp. 179–196, 2002.
- [25] M. Sattinger, “Queueing and searching,” Ph.D. dissertation, Dept. Econ., Univ. Albany, Albany, NY, USA, 2010.
- [26] M. Coles and E. Smith, “Marketplaces and matching,” *Int. Econ. Rev.*, vol. 39, no. 1, pp. 239–254, 1998.
- [27] E. Ebrahimi and R. Shimer, “Stock-flow matching,” *J. Econ. Theory*, vol. 145, no. 4, pp. 1325–1353, 2010.
- [28] P. Á. de Toledo, F. Núñez, and C. Usabiaga, “La función de emparejamiento en el mercado de trabajo español,” *Revista de Economía Aplicada*, vol. 16, no. 3, pp. 5–35, 2008.
- [29] P. Á. de Toledo, F. Núñez, and C. Usabiaga, “An empirical analysis of the matching process in the Spanish public employment agencies: The vacancies,” Dept. Econ., Universidad Pablo de Olavide, Seville, Spain, Working Papers 11.03, 2011.
- [30] R. Cotterman and F. Peracchi, “Classification and aggregation: An application to industrial classification in CPS data,” *J. Appl. Econometrics*, vol. 7, no. 1, pp. 31–51, 1992.
- [31] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.
- [32] P. Á. de Toledo, F. Núñez, and C. Usabiaga, “An empirical approach on labour segmentation. Applications with individual duration data,” *Econ. Model.*, vol. 36, pp. 252–267, Jan. 2014.
- [33] P. Á. de Toledo, F. Núñez, and C. Usabiaga, “Matching and clustering in square contingency tables. Who matches with whom in the Spanish labour market,” *Comput. Statist. Data Anal.*, vol. 127, pp. 135–159, Nov. 2018.
- [34] D. Power and M. Lundmark, “Working through knowledge pools: Labour market dynamics, the transference of knowledge and ideas, and industrial clusters,” *Urban Stud.*, vol. 41, nos. 5–6, pp. 1025–1044, 2004.
- [35] R. Eriksson and U. Lindgren, “Localized mobility clusters: Impacts of labour market externalities on firm performance,” *J. Econ. Geography*, vol. 9, no. 1, pp. 33–53, 2009.
- [36] O. Figueiredo, P. Guimarães, and D. Woodward, “Firm–worker matching in industrial clusters,” *J. Econ. Geography*, vol. 14, no. 1, pp. 1–19, 2014.

- [37] J. P. Sloane, D. P. Murphy, I. Theodossiou, and M. White, "Labour market segmentation: A local labour market analysis using alternative approaches," *Appl. Econ.*, vol. 25, no. 5, pp. 569–581, 1993.
- [38] A. Chakraborty, M. A. Beamonte, A. E. Gelfand, M. P. Alonso, P. Gargallo, and M. Salvador, "Spatial interaction models with individual-level data for explaining labor flows and developing local labor markets," *Comput. Stat. Data Anal.*, vol. 58, pp. 292–307, Feb. 2013.
- [39] H. Sala and P. Trivín, "Labour market dynamics in Spanish regions: Evaluating asymmetries in troublesome times," *Series*, vol. 5, no. 2, pp. 197–221, 2014.
- [40] H. S. Zwick and S. A. S. Syed, "The polarization impact of the crisis on the Eurozone labour markets: A hierarchical cluster analysis," *Appl. Econ. Lett.*, vol. 24, no. 7, pp. 472–476, 2017.
- [41] U. Blien and F. Hirschenauer, "A new classification of regional labour markets in Germany," *Let. Spatial Resour. Sci.*, vol. 11, no. 1, pp. 17–26, 2018.
- [42] M. de Trabajo and Migraciones y Seguridad Social, "La Muestra Continua de Vidas Laborales," Estadísticas, Presupuestos y Estudios, Seguridad Social, Madrid, Spain, Tech. Rep., 2016.
- [43] A. Spark. (2018). *Clustering—Spark 2.3.0 Documentation*. Accessed: Jul. 12, 2018. [Online]. Available: <https://spark.apache.org/docs/2.3.0/ml-clustering.html>
- [44] Clustering Stata. (2018). *Cluster Analysis—Stata*. Accessed: Jul. 12, 2018. [Online]. Available: <https://www.stata.com/features/cluster-analysis/>
- [45] P. Yildirim and D. Birant, "K-linkage: A new agglomerative approach for hierarchical clustering," *Adv. Electr. Comput. Eng.*, vol. 17, no. 4, pp. 77–88, 2017.
- [46] X. Wu, T. Ma, J. Cao, Y. Tian, and A. Alabdulkarim, "A comparative study of clustering ensemble algorithms," *Comput. Electr. Eng.*, vol. 68, pp. 603–615, May 2018.
- [47] R. Pérez-Chacón, J. M. Luna-Romera, A. Troncoso, F. Martínez-Álvarez, and J. C. Riquelme, "Big data analytics for discovering electricity consumption patterns in smart cities," in *Energies*, vol. 11, no. 3. Basel, Switzerland: MDPI, 2018.
- [48] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
- [49] J. M. Luna-Romera, M. Martínez-Ballesteros, J. García-Gutiérrez, and J. C. Riquelme, "External clustering validity index based on chi-squared statistical test," *Inf. Sci.*, vol. 487, pp. 1–17, Jun. 2019.



interests include labor economics and the economy of the electricity sector.



MARÍA MARTÍNEZ-BALLESTEROS received the M.Sc. degree in computer engineering and the Ph.D. degree in computer science from the University of Seville, Spain, in 2012. Since 2009, she has been with the Department of Computer Science, University of Seville, where she is currently an Associate Professor. Her primary research interests include data mining, machine learning techniques, association rules, evolutionary computation, and big data.



JOSÉ C. RIQUELME received the M.Sc. degree in mathematics and the Ph.D. degree in computer science from the University of Seville, Spain. Since 1987, he has been with the Department of Computer Science, University of Seville, where he is currently a Full Professor. His primary research interests include data mining, machine learning techniques, and evolutionary computation.



interests include macroeconomics and labor economics.



JOSÉ MARÍA LUNA-ROMERA received the M.Sc. degree in software engineering and technology, in 2012, and published his master's thesis on data mining applied to earthquakes prediction. He has been a Ph.D. Research Student with the University of Sevilla, Spain, since November 2015, after being awarded a four-year research scholarship by the Spanish Government. His current research interests include clustering analysis and more generally data mining and big data.

...