Are Labour Markets Segmented in Developing Economies? A Clustering Approach for Colombian Workers*

David Rodríguez Guerrero**
Universidad Externado de Colombia, Bogotá

Jorge Eliecer Quintero***
Universidad Externado de Colombia, Bogotá

Abstract

Labour markets in developing economies are usually thought to be segmented. Differences in productivity, red tape, and high taxes create a divide between a modern and an excluded traditional sector. More recently, some scholars have challenged this view. In this article, we propose to test the segmented markets hypothesis using a clustering method applied to Colombian workers. Following Anderson et al. (1987) we hypothesize that if the first view prevails, the labour market has well-defined worker clusters that our empirical strategy could uncover. Using the FAMD-K-means algorithm we find three clusters: one comprises half the workforce, has workers with secondary education or vocational training, without labour contracts, and median earnings slightly above the minimum wage. The second group comprises 37% of the workforce, older workers with even lower earnings and educational achievement, with more precarious jobs. The last cluster comprises good quality jobs, mostly with indefinite labour contracts, with workers with university degrees and median earnings close to four times the minimum wage. We statistically tested the differences between the informality definition and our method and found that the traditional measures have an important correlation with the clusters resulting from our model.

Keywords: Labour informality; clustering methods; unsupervised machine learning; segmented market hypothesis.

JEL: J21; J23; J42; J82; C38.

** Lecturer and Researcher at Universidad Externado de Colombia, Faculty of Economics (Bogotá, Colombia). E-mail: david.rodriguez@uexternado.edu.co iD https://orcid.org/0000-0003-2670-3786 corresponding author

*** Research Assistant at Universidad Externado de Colombia, Faculty of Economics (Bogotá, Colombia). E-mail: jorge.quintero46@est.uexternado.edu.co iD https://orcid.org/0009-0003-6512-2025

¿Son los mercados laborales segmentados en los países en desarrollo? una aproximación de clustering de los trabajadores colombianos

Resumen

Los mercados laborales en las economías en desarrollo suelen considerarse segmentados. Diferencias en productividad, la burocracia y los impuestos elevados crean una brecha entre un sector moderno y otro tradicional y excluido. Más recientemente, algunos académicos han desafiado esta perspectiva. En este artículo proponemos poner a prueba la hipótesis de mercados segmentados mediante un método de clustering aplicado a los trabajadores colombianos. Siguiendo a Anderson et al. (1987), sugerimos que, si prevalece la primera perspectiva, el mercado laboral tiene grupos de trabajadores bien definidos que nuestra estrategia empirica puede encontrar. Usando el algoritmo FAMD-K-means encontramos tres clusters: uno con la mitad de la fuerza laboral, con trabajadores con secundaria o formación profesional, sin contratos laborales y con ingresos medios ligeramente superiores al salario mínimo. El segundo grupo comprende el 37% de la fuerza laboral y está compuesto por trabajadores mayores con ingresos y logros educativos aún más bajos y empleos más precarios. El último grupo comprende empleos de buena calidad, en su mayoría con contratos laborales indefinidos, para trabajadores con títulos universitarios e ingresos medios cercanos a cuatro veces el salario mínimo. Testeamos estadísticamente diferencias entre la definición de informalidad y nuestro método, y encontramos que las medidas tradicionales tienen una correlación importante con los grupos encontrados.

Palabras clave: informalidad laboral; métodos de clustering; algoritmos de aprendizaje no supervisado; hipotesis de mercados laborales segmentados.

Os mercados de trabalho são segmentados nos países em desenvolvimento? Uma abordagem de agrupamento para trabalhadores colombianos

Resumo

Os mercados de trabalho nas economias em desenvolvimento são geralmente considerados segmentados. As diferenças de produtividade, a burocracia e os altos impostos criam uma lacuna entre um setor moderno e um setor tradicional e excluído. Mais recentemente, alguns acadêmicos questionaram essa perspectiva. Neste artigo, propomos testar a hipótese dos mercados segmentados usando um método de agrupamento aplicado aos trabalhadores colombianos. Seguindo Anderson et al. (1987), sugerimos que, se a primeira perspectiva prevalecer, o mercado de trabalho tem grupos bem definidos de trabalhadores que nossa estratégia empírica pode encontrar. Usando o algoritmo FAMD-K-means, encontramos três grupos: um com metade da força de trabalho, com trabalhadores com ensino médio ou treinamento vocacional, sem contratos de trabalho e com ganhos médios ligeiramente acima do salário mínimo. O segundo grupo compreende 37% da força de trabalho e é composto por trabalhadores mais velhos, com rendimentos e nível de escolaridade ainda mais baixos e empregos mais precários. O último grupo compreende empregos de boa qualidade, em sua maioria com contratos de trabalho permanentes, para trabalhadores com diploma universitário e ganhos médios próximos a quatro vezes o salário mínimo. Testamos estatisticamente as diferenças entre a definição de informalidade e nosso método, e descobrimos que as medidas tradicionais se correlacionam significativamente com os grupos encontrados.

Palavras-chave: informalidade do trabalho; métodos de agrupamento; algoritmos de aprendizado não supervisionado; hipótese de mercado de trabalho segmentado.

Introduction

In developing economies, labour income represents the main resource for families to overcome poverty. This is especially true if we consider the lack of a safety net and the concentration of non-labour income at the top the income distribution. However, labour markets in most of these countries are characterised by the apparent coexistence of two very dissimilar productive sectors, a high productivity, modern sector able to comply with regulations, and a low productivity, traditional sector apparently excluded. In most cases, the low-productivity sector employs a high share of workers, and therefore, family labour incomes are insufficient for a decent life for a high proportion of the population.

This divide could be considered both cause and consequence of underdevelopment. On the one hand, the size of the traditional sector implies that governments are unable to collect enough revenues, providing fewer public goods such as education or infrastructure resulting in underdevelopment. On the other hand, the existence of two sectors reflects the inability of underdeveloped economies to increase productivity for all workers and firms.

From a measurement point of view, economists have usually named the traditional sector the informal sector. Originally, the "Informality" expression referred to small-scale economic activities hidden from government supervision; often denoted as the "underground", "unrecorded", "non-protected" or "grey" sector of the economy (ILO, 1972). More recently, the definition moved away from underground and illegal activities. Its measurement, at least in relation to the segmentation of labour markets has been mostly confined to two views: productivity —establishment size— or legalistic —contribution to social security—. However, the two measures are not observationally equivalent (Henley et al., 2009) and additionally, further classifications such as "informal employment outside the informal sector have been proposed" (Hussmanns, 2004) implying a lack of intragroup homogeneity. In what follows, we depart from these two *ad-hoc* measures of labour market segments.

From a theoretical point of view, segmentation has been usually analysed within two perspectives: exclusion and exit (Perry et al., 2007). The exclusion point of view assumes that complying with regulations is expensive for small companies and workers whose productivity is low relative to the burden of regulation and taxes; therefore, these firms and workers are excluded from the so-called modern sector (de Soto, 1989). The exit suggests that firms and workers choose whether to contribute to taxes and social insurance —or not—, having both options available. In the cost-benefit analysis, some of them decide not to comply considering weak law enforcement and the availability of subsidised social protection (Maloney, 1999). Whether economies are closer to one view or the other matters for policy proposals and evidence is not sided with any of them. If the first view is closer to reality, governments must pursue policies that increase productivity, if the second view is more realistic, increasing compliance is a must.

In this article, we test the segmented market hypothesis by means of an unsupervised machine-learning algorithm. Following Anderson et al. (1987) we argue that if this hypothesis

prevails, the labour market has segments with reduced intragroup and high extra group heterogeneity implying an important divide between groups of workers in the labour market. If the labour market is not segmented, there should not be important differences between the segments arising from our clustering algorithm, there is no disadvantaged or excluded sector and we must consider the labour market in developing countries as one for policy analysis. Additionally, we compare the *ad-hoc* classification of workers with the informality definitions and our clustering approach.

Our objective is two-fold. First, we would like to analyse the Colombian labour market through clustering, a machine-learning technique highly unexplored in the economics literature for the country. We devote our efforts to present the methodology in a concise way to non-data scientist. Secondly, clustering provides new insights into the informality-formality divide, and more importantly, it does not require an a priori labelling of workers. This is relevant as labour informality measurement moves between the legalistic and productivity views and the use of one or the other is highly debated.

We find an important divide between the segments resulting from our clustering algorithm, and more importantly, the divide that the data shows is correlated with the traditional definitions of informality. More precisely, we found that the best strategy to analyse the Colombian labour market is a combination of Factor Analysis of Mixed Data (FAMD) as a pre-processing technique that transforms categorical variables to numerical ones and K-Means as the clustering algorithm. We also find that the optimal number of clusters with this algorithm is three and that there are important differences between the workers in each cluster, especially on the educational and earnings dimensions. Comparing the categories resulting from the clustering algorithm with traditional measures of labour informality, we reject the null hypothesis of independence between the two classifiers. This evidence points to the presence of segmentation in the Colombian labour market.

The article is divided into five sections with this introduction being the first one. In the second section, we review recent articles that test the segmented market hypothesis and literature on the use of clustering analysis with labour market data. The third section describes the clustering methods and the data for the exercise. The fourth section presents the clustering results, its comparison to the traditional informality measurement, and the statistical testing of the segmented market hypothesis. The fifth section concludes.

Testing the labour market segmentation hypothesis: a review

In this section, we first review traditional economic literature that tests the segmented market hypothesis. Next, we review the clustering approach to analyse whether labour markets are segmented or not. The discussion stresses the reasons behind one view or the other and the empirical strategies employed to test the hypothesis.

Traditional econometric models that test segmentation[1]

In its origins, labour market segmentation in developing countries has been considered the result of high labour market regulations and low productivity for firms and workers. The first economic analyses of segmentation are based on the two-sector model of Harris and Todaro (1970) in which rural workers move to the urban labour market guided by important differences in wages between sectors. More recent literature that focuses on informality highlights that workers and firms analyse expected returns and costs of choosing the formal or the informal sector, in top of weak low enforcement, agents consider taxes, social insurance payments, the availability of non-contributory social security or conditional cash transfers programmes in case of working as informal (Maloney, 1999)[2]. If firms or workers are deciding not to comply it is because of an expected comparative advantage, but in practice, they are free to move between sectors. The high movement of workers between informality and formality found in countries such as Mexico is usually presented as validation of this view.

In this regard, Farné (1990) indicates that it is possible that segmentation is not fixed characteristic of the labour market, but instead, that the movement between sectors could depend on the business cycle. He uses household survey data from Colombia and Mincer equations for formal and informal workers to demonstrate that in expansions, the segmentation of labour markets is reduced as the dispersion of earnings increases but the earnings differential between sectors is reduced. The opposite is true for recessions.

Magnac (1991) and Pradhan and van Soest (1995; 1997) test the hypothesis of labour market segmentation in Latin America using microeconomic data. The first one uses data for Colombia and compares bivariate Tobit estimates of two types of models, a "free movement between" sectors model and a model where workers "queue for a formal job". Magnac (1991) finds that the hypothesis of a competitive equilibrium that is free movement is not rejected. Pradhan and van Soest (1995; 1997) study labour formality choice in Bolivia. In the first paper they use ordered probit models (i.e., informal sector is inferior for workers, implying segmentation) and multinomial logits (there is no ordering of sectors). They conclude that multinomial logits better characterise women's choices and ordered probits are better suited for modelling men's choices (segmentation). For the second paper, the authors propose a structural labour supply model in which couples choose sector and hours of work based on sector-specific wages. Simulations indicate that a 10% decrease in formal sector wages moves 2.1% of male workers from the formal to the informal sector and increases female participation by 0.4%.

Using panel data for Mexico, Maloney (1999) and Gong et al. (2004) use multinomial logit models to explore transitions between formality and informality and the worker´s characteristics

---

1   This section draws heavily on Rodríguez (2021).
2   See also Neffa (2008) and Neffa (2023) for a detailed overview of theories of labour market segmentation.

that shape them. Maloney (1999) uses a model for three sectors: self-employed, informal employees, and, formal employees suggesting that overall, there is a high degree of mobility between sectors and that the length of job tenure is similar for all sectors, which implies that is not the case that workers arrive in the formal sector and stay there forever, but they also move between sectors rejecting the segmentation hypothesis. On the other hand, informal workers do not seem to be queuing for formal jobs because higher experience is not observed to determine their transition to formality. Gong et al. (2004) use a dynamic model with random effects and divide the working age population into three categories: non-working, formal and informal. As in Maloney (1999) they find that movement between these states is considerably high, secondly, that the barriers to the formal sector are higher than for the informal sector for lower educated individuals, thirdly, a strong state persistence for educated individuals.

In a more recent study, Rodríguez (2021) proposes the estimation of a structural labour supply model with job availability restrictions for Colombia; the model is based on the RURO model developed by Aaberge et al. (1995). After some pro formality policy simulations such as increases in educational attainment and reduced social insurance payments, results indicate that job availability in the formal sector does not increase substantially for informal workers validating the segmented market hypothesis.

Concluding, there seems to be no definitive answer to the validity of the segmented market hypo-thesis in developing countries in Latin America employing traditional econometric methods.

Clustering analysis of labour markets

Clustering is a standard method of unsupervised learning employed to join or segment a collection of objects in subsets or clusters, this is done in such a way that the objects in a specific segment share similar features among them, but they substantially differ with the objects in other groups (Hastie et al., 2009). This implies that an object belonging to a cluster reflect the most important sources of differences (or heterogeneities) among a dataset (Martin & Okolo, 2022).

It is common in the labour economics literature to find clustering analysis to test several theories, especially those related to the segmented market hypothesis (Anderson et al.,1987; Sousa-Poza, 2004), or to analyse if labour market heterogeneity is guided by productivity or workers characteristics (Martin & Okolo, 2022). In the first case Anderson et al. (1987) develop a segmentation model for the US labour market using the PSID data. They test the dual labour market hypothesis by creating labour features indices to classify jobs as good or bad. They use an agglomerative clustering method and find no evidence of a dual (or multiple) labour market in the US. Sousa-Poza (2004) also analyses the segmented market hypothesis for another developed economy, Switzerland, he uses labour surveys and proposes three metho-dological strategies: a clustering analysis, a switching model, and a bivariate probit model with endogenous selection. For the clustering analysis, he uses a Hierarchical Clustering with the

"Average linkage" as the distance measure. He analyses if jobs (defined as pairs of industries and occupations) could be placed in segments that depend on worker characteristics such as age, gender, education or on the job training. The author identifies seven clusters, but only one of considerable size and argues that there is no clear evidence of segmentation.

On the other hand, Martin & Okolo (2022) use data from the UK's labour market and a K-Medoids algorithm for binary variables to validate whether UK's labour market heterogeneity is exclusively due to differences in productivity (measured by worker's education or occupation) instead of other worker's characteristics such as gender, ethnicity etc. They find that the UK's labour market is composed by two segments, but suggest that the labour market is not segmented because not all workers in the high productivity cluster have a high education level, but some low productivity workers have high education levels.

Clustering of labour markets in developing economies

López-Roldan and Fachelli (2021) study the Spanish and Argentinian labour markets, they hypothesise that there is not a unique market that adjusts labour supply and demand. The authors use a set of demand and supply variables for both countries and use an agglomerative hierarchical clustering algorithm and the Ward method (i.e. with clustering optimization using mobile centroids). They find that the labour markets of the two countries could be represented by four segments. The first one captures those workers with precarious labour conditions, without labour contracts, informal or casual workers (part time) with low wages. This segment is constituted mainly by women and young workers with low education levels. The second cluster is similar to the first one but has many more foreign workers and a more balanced gender composition. The third cluster is characterized by male adults with secondary education or vocational training with full time and stable contracts, mainly working in technical or administrative positions, with an average to high labour income. The last cluster comprises those workers with the best labour conditions and highest education levels (mainly professionals) with supervising roles, working in large companies, and dominated by female adults.

Lastly, Howell (2011), examines the segmented market hypothesis for the city of Urumqi in the Xinjiang Uyghur Autonomous region in China, with a special focus on internal migration towards the region and its ethnical diversity. The author combines Principal Components Analysis (PCA) and Hierarchical Clustering/ K-means Clustering and employs as features for the algorithm: ethnicity, migratory status, employment type, occupation, and industry. He finds three clusters: the first one has high education and high earnings and is mainly composed by independent workers. The second cluster has mostly employees with lower earnings than the first cluster. The last cluster is comprised of women, migrants not from the Han ethnic group, and other workers with low education levels and earnings. The author concludes the segments are highly correlated to migration and ethnicity.

Empirical strategy and data

In this section, we start by presenting the data for the Colombian labour market; next, we describe the pre-processing of the data and give a brief description of the three clustering algorithms proposed. Lastly, we present some validation measures and a statistical test to analyse whether the traditional informality definition is correlated to the clusters we find in the data.

Data

For the purposes of our analysis, we focus on a developing country: Colombia. We use the main labour household survey for this country: the Great Integrated Household Survey (GEIH) for the year 2019. GEIH is a rich survey with detailed information on employment, incomes, as well as household and personal characteristics such as education, gender, or age but also on job features such as sector, occupation or working hours. The sample for 2019 is comprised of 756063 observations, with 316562 of them being workers.

Methodology

In data science projects, it is a good practice to follow a series of pre-processing steps before the model is deployed; this is especially true for clustering analysis where the quality of the pre-processing will determine the homogeneity of the data in each cluster and the heterogeneity between clusters. First, any machine-learning algorithm requires a delicate process of variable selection based on the knowledge the researcher has of the problem. Secondly, the researcher must understand the nature of the data, that is whether the variables are numerical (continuous or discrete) or categorical (binary, categorical or ordinal) or if the dataset has a combination of both types. The second step will determine the transformation required for the algorithm. For instance, if the variable is categorical and requires an encoding to make it numeric, or if the variable is numerical and the algorithm requires categorical data, a data binning is required. If the variable is continuous but there are some other continuous variables a re-scaling of variables is usually needed: standardisation or normalisation. Thirdly, in the case of clustering, a similarity/dissimilarity measure is needed in order to create segments of data with high intragroup homogeneity and extra group heterogeneity; this of course will determine the clustering algorithm to be used. Lastly, in the case of clustering, a cost/benefit measure should indicate the optimal number of clusters.[3]

In the case of clustering, the great challenge is how to deal with different types of data at the moment of creating the segments. It is well known that most clustering algorithms admit only one type of variables. For instance, in the case of numerical variables we found the algorithms

---

3   Methodologies such as the "Cross-Industry Standard Process for Data Mining (CRISP-DM)" are a good point of reference for the development of data science projects.

Hierarchical Clustering, K-Means, Fuzzy K-Means, or Probabilistic Distance Clustering. In the case of categorical variables K-Modes, Fuzzy K-Modes, etc.[4]. There are very few clustering algorithms for mixed datasets, being the most known the K-Prototypes.

To deal with mixed data, several approaches have been proposed. Following van de Velden, et al. (2019) there are three of them:

i.   Basic pre-processing to transform all variables to the same type and later use of one algorithm of those presented above.
ii.   Create a similarity/dissimilarity measure for mixed data, for instance Gower distance (Gower, 1971), and use a distance clustering method such as Partitioning Around Medoids (PAM) (Apitzsch & Ryeng, 2020). Other techniques use extensions of K-Means for mixed data such as the K-prototypes algorithm (Akay & Yüksel, 2018), 2018) K-Means to Mixed Data (Koren, et al., 2019), K-Harmonic Means (Ahmad & Hashmi, 2016) and Modha–Spangler Convex K-Means Clustering (Modha & Spangler, 2003).
iii. Use dimensionality reduction techniques and clustering, for instance using Tandem analysis, which consists of applying techniques such as Factor Analysis of Mixed Data (FAMD or PCAMIX) and later use the resulting components to apply a distance-based clustering algorithm such as K-means.

Variable selection

For the variable selection, we took as reference previous segmentation studies applied to the labour market such as Anderson et al. (1987), Boston (1990), Gittleman and Howell (1995), López-Roldán and Fachelli (2021) and Martin and Okolo (2022). One thing in common among these studies is the use of variables related to both labour supply and demand. Among the demand dimension we find labour conditions for workers in aspects such as stability, required qualifications for the job, earnings, and other firm characteristics. From the supply side we find socio-demographic variables such as gender, age, ethnicity, nationality, or education level.

The set of variables employed in this article could be found at Table 1 distinguishing by supply or demand variables. There are 13 variables with two of them being numerical and 11 categorical.

---

4   A full review of clustering algorithms can be found at in Aggarwal and Reddy (2014) or in Hennig et al. (2015).

Table 1. Dimensions and variables for the segmentation model

Dimension	Indicator/ Variable	Categories / detail	Data type
a. Labour Demand			
1. Stability	Contract type and duration	Indefinite, fixed ≤ 6 months, fixed > 6 months, other	Nominal
	Working time	<30 hrs (part-time), 30 a 50 hrs (full-time), >50 hrs (extra-time).	Nominal
	Tenure at the current firm	< 1 year, 2 to 3 years, 4 to 10 years, 11 to 20 years, > 20 years	Nominal
2. Qualification	Occupation	* Senior officials and managers * Professionals * Technicians and associate professionals * Clerks * Service and sales workers * Skilled agricultural * Craft and trades workers * Plant and machine operators * Elementary occupations	Nominal
	Type of work	* Domestic worker * Day laborer or peon * Worker or employee of a private company * Government worker or employee * Employer * Self-employed * Other work	Nominal
3. Earnings	Labour income	COP monthly	Numeric
4. Company's characteristics	Industry	* Agriculture and Fishing * Mining, Manufacturing and Utilities * Construction * Wholesale and retail trade * Hotels and restaurants * Transport and communication * Financial intermediation * Real estate and business activities * Public administration and defence * Education * Health and social work * Other	Nominal
b. Labour supply			
5. Gender	Gender	Male or Female	Binary
6. Age	Age	Age in years	Numeric
7. Ethnicity	Ethnicity	* Black, mulato, afrocolombian * Indigenous * Raizal from San Andres, Providencia, Santa Catalina * Palenquero from San Basilio * Romani	Nominal
8. Nationality	Nationality	Native or Foreigner	Binary
9. Education	Education level	Without educ, Primary, Secondary, middle school, technical of technological ed, Undergraduate and Postgraduate.	Ordinal
	Graduated	Yes (it has a diploma including high-school diploma) or not	Binary
	Working and studying	Yes, or not	Binary

Notes: (1) National classification of occupations (CNO-70) National Training Service (SENA). (2) GEIH 2019 uses International Standard Industrial Classification of all Economic Activities (ISIC Rev 3).
Source: Own elaboration based on other authors and using GEIH (DANE, 2019).

Data pre-processing

For our research we use three distance-based clustering algorithms: Tandem Analysis FAMD combined K-Means Clustering (in what follows Tandem: FAMD-K-means) K-Modes and K-Prototypes. As discussed before, each algorithm requires specific transformations before its deployment.

Considering the nature of our data, we carry out the following steps before model estimation: in the case of the K-Prototypes algorithm, the numerical variables (age and earnings) are scaled using the min-max method to place them in the interval zero to one. This is done to avoid that some features(variables) dominate the clustering process. In the case of FAMD-K-Means we transform categorical variables to dummy variables (one-hot encoding). Each dummy variable is divided by the square root of the proportion of observations in the associated category (dummy=1). Lastly, numerical variables are standardized. In the case of the K-Modes algorithm we discretised age (at 5 years intervals) and earnings (as deciles) to have only categorical variables in the dataset.

Clustering algorithms used

Being one of the main machine-learning tools, clustering nowadays has a great variety of algorithms. These algorithms could be classified in two types: top-down (or divisive) and bottom-up (or agglomerative). In the first group, we find the algorithms K-means, K-Medoids among others for which we start with the entire dataset and the objective is to divide it according to similarity measures known as distances. In the second group within the Hierarchical Clustering family, we find the hierarchical agglomerative clustering (HAC) algorithm, where each cluster is built starting by one item per cluster, and later joining similar clusters trying at each step to minimize the intragroup variance of the joining clusters.

Moreover, the literature recognises that algorithm and distance measure selection depend on the nature of the data, the research objectives, the sample size, and computational power. As it was mentioned before, this study uses three clustering algorithms FAMD-K-Means, K-Modes, and K-Prototypes. Considering the large sample size, the selected top-down algorithms substantially reduce computational times relative to other available algorithms such as hierarchical clustering for mixed data.[5] We now present the specificities of the three algorithms.

Tandem: FAMD-K-Means

As it is mentioned by van der Velden, et al. (2019) FAMD-K-Means could be described as a clustering technic that combines dimensionality reduction and a clustering algorithm such as K-Means (Bock, 1987). Considering the nature of our data, in this study we use Factor Analysis of Mixed Data (FAMD) as the dimensionality reduction technique. To deploy FAMD-K-Means categorical and numerical variables must be transformed in such a way that the influence of

---

5   As it is mentioned by Grané and Sow-Barry (2021) hierarchical clustering could also work with mixed data using the Gower coefficient. However, it is known that with large samples, the algorithm is computationally expensive.

each variable is balanced, that is that both types of variables are on equal foot to determine the dimensions or principal components of the first part of the algorithm.

Regarding the FAMD algorithm, it could be thought as a PCA applied to the numerical variables and a Multiple Correspondence Analysis (MCA) applied to the categorical data. The algorithm delivers a group of principal components, which are numerical linear combinations of the mixed-type input data. These variables are in turn used in the second step of the FAMD-K-Means algorithm, a K-Means clustering for the transformed data. Following Huang (1998) K-Means could be described as follows: suppose a set of numerical variables X with n observations, values, or objects in each one. An integer number of clusters needs to be formed minimising the sum of the squared errors (also called inertia)[6] between each object and the centroid of its cluster S. The centroid is typically an object representing the average of the values of the objects in the cluster while the error is defined as the Euclidian distance, more formally, the optimization problem is:

$$\min_{\mu_i} = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \qquad (1)$$

Where x_j is an observation or object that belongs to cluster S_i, μ_i is the centroid of cluster S_i and $||x_j - \mu_i||^2$ is the square of the Euclidean distance between x_j and the centroid of the cluster. The objective is to iteratively define a new μ_i and redefine the cluster to which each object belongs based on the minimum Euclidean distance between the object and each centroid. Following Pedregosa et al. (2011) the K-Means algorithm performs the following steps:

1. At the first step k centroids are defined. The centroids could be k observations picked at random (AKA Forgy method) or randomly assigning a cluster to each observation and computing the centroid using the mean of the object in the cluster. (AKA random partition)
2. Assignment step: assign each object to the cluster for which the distance to its centroid is the lowest.
3. Update step: recalculate de centroid of the cluster as the mean of the objects in the cluster.
4. Repeat steps 2 and 3 until there is no further improvement in equation 1 and the assignment of objects to each cluster. Given the nature of the optimization problem there is no guarantee that there is an optimal solution.

K-Modes

The K-Modes algorithm developed by Huang (1997b) is an extension of K-Means but for categorical data. In a similar spirit of K-Means, K-Modes calculates distances but instead of using the Euclidean distance it uses the following measure of dissimilarity (Hamming distance):

---

6   Following Pedregosa et al. (2011), inertia could be described as a measure of the degree of internal coherence between clusters.

Suppose two objects or observations X and Y with m categorical attributes or variables, a dissimilarity measure is defined as the sum of mismatches between the two objects in Equation 2, where a mismatch in the j attribute could be defined as in Equation 3. The fewer the number of mismatches the higher the similarity between the objects.

$$d(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad (2)$$

Where:

$$\delta(x_j, y_j) = \begin{cases} 0 & \text{if } x_j = y_j \\ 1 & \text{if } x_j \neq y_j \end{cases} \quad (3)$$

Notice that in this case the algorithm gives the same weight to each attribute. We could also modify the dissimilarity measure by considering the frequency of each category in the dataset as follows:

$$d(X, Y) = \sum_{j=1}^{m} \frac{(n_{x_j} + n_{y_j})}{n_{x_j} n_{y_j}} \delta(x_j, y_j) \quad (4)$$

Where n_{xj} and n_{yj} are the number of objects in the dataset containing x_j and y_j respectively, for attribute j. As described by Huang (1997b) the K-Modes algorithm performs the following steps:

1. Select k initial modes, one for each cluster, either selecting k random objects from the dataset or picking the objects in the dataset closer to the most frequent categories in the categorical variables in the dataset.
2. Assign each object to the nearest cluster, that is the cluster for which the distance d is the lowest.
3. Update the modes of each cluster.
4. Repeat step 3 until no object changes its cluster.

K-Prototypes

The K-prototypes algorithm, developed by Huang (1997a), is one of the most used clustering methods as a result of the good segmentation properties displayed (Preud'homme, et al., 2021). It is a variant of the K-means algorithm but considers a dissimilarity measure that consists of a Euclidean distance for numerical data and a frequency distance-based metric for categorical data. In that sense, it is ideal to cluster mixed type datasets.

Following Huang (1997a) Grané & Sow-Barry (2021) and Preud'homme, et al., (2021), the algorithm defines prototypes as the group centroids. They are built based on the average of the numerical variables and the modes of the categorical variables in the cluster.

The distance between to objects X and Y each one with p attributes is defined as follows:

$$d(X,Y) = \sum_{j=1}^{q}(x_j - y_j)^2 \; + \; \gamma \sum_{j=q+1}^{p} \delta(x_j, y_j) \qquad (5)$$

Where the first term is the squared Euclidian distance for numerical variables 1 to q and the second is the measure of dissimilarity for categorical objects $q+1$ to p, as presented in equation 2. The weight γ is used to avoid favouring one information type (numerical or categorical) over the other. It could be supplied by the user or estimated using a combined variance of the data.

Besides determining the number of cluster (k), the algorithm requires a centroid for each cluster. For this part, there are two commonly used methods. The first one was developed by Huang (1997b) and randomly selects k different objects as centroids. The second one follows Cao, et al., (2009) and uses the density of each attribute placing the centroids iteratively at sparse points with high densities of the attributes.

Once the centroids are initialized the algorithm follows similar steps updating the objects in each cluster and the centroids as K-Modes or K-Means. The process continues until each cluster is stable.

Determining the number of optimal clusters

One of the main challenges when working with top-down clustering methods is determining the optimal number of clusters. There are two alternatives to choose this number: firstly, the researcher could resort to previous literature and alternatively, an algorithm could be used. Regarding the first alternative, the literature review from Section 2 indicates that for labour market segmentation, the optimal number of clusters is around three or four. On the other

hand, considering the mixed type of nature of our data, there are two common methodologies to determine the number of clusters: the so-called Elbow Method and the Silhouette Analysis. In this research, we use the first method, because as highlighted by the simulation exercises of Aschenbruck and Szepannek (2020) and Grané and Sow-Barry (2021), Silhouette Analysis is computationally expensive for large datasets and presents good features mostly for small datasets (400 observations or less).

The Elbow method relates the value of the cost function to different numbers of clusters, with the cost function depending on the clustering algorithm used as described below.

In the case of the K-Means algorithm we define the cost, also called inertia as the sum of the squared distances between each object and the centroid of its cluster:

$$\text{Inertia} = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \qquad (6)$$

In the case of K-Modes the cost function is the sum of the hamming distances (as defined in Equation 2) of each object and its cluster:

$$\text{cost} = \sum_{l=1}^{k} \sum_{i=1}^{n} d(X_i, C) \qquad (7)$$

In the case of K-Prototypes, for a specific cluster l, the cost function is given by:

$$E_{kproto} = \sum_{l=1}^{k} (E_l^r + E_l^c) = \sum_{l=1}^{k} E_l^r + \sum_{l=1}^{k} E_l^c = E^c + E^r \qquad (8)$$

Where, y is the total cost of the numerical and categorical data in cluster , respectively.

If the number of clusters increases, the average cost is reduced. The value for the optimal number of clusters results from the point in which the cost function decreases the most and is known as the Elbow or inflection point, increasing the number of clusters behind this point won't make significant gains in cost reduction.

Validating the resulting clusters

Following Aggarwal and Reddy (2014), and Apitzsch and Ryeng (2020), cluster validation is required to avoid finding spurious patterns in the data, and therefore, to be able to conclude that the resulting clusters really represents true sub-groups of the data. Moreover, cluster validation also is an adequate tool to compare the results of several clustering algorithms. As it is mentioned by Aggarwal and Reddy (2014) there is no consistent protocol to validate clusters. However, there are usually two validation procedures: internal validation, and external validation. The main difference between the two is that external validation requires external information to check the resulting clusters[7].

The internal validation instead evaluates the goodness of a clustering structure without resorting to external labels for the objects. Some of the most used methods for internal validation are the Silhouette index (S), the Calinski-Harabasz index (CH) or Variance Ratio Criterion and the Davies-Bouldin index (DB).

The Silhouette index (S), developed by Rousseeuw (1987), evaluates the goodness of the clustering algorithm using two dimensions: the average distance between an object (and all the other objects in its cluster $(a(x_i))$, and the average distance of the object and all the elements of the nearest cluster $b(x_i)$:

$$s(x_i) \;=\; \frac{b(x_i) - a(x_i)}{\max\big(a(x_i), b(x_i)\big)} \quad \text{(9)}$$

Where, $s(x_i)$ represents the silhouette width of object x_i and $b(x_i) = \min\{d\_l\,(x_i)\}$ with $d_l(x_i)$ being the average distance between x_i and objects in cluster l, with $l \neq k$. The average distance of the entire dataset is the average of $s(x_i)$ for all x_i objects. By construction, the silhouette width belongs to the interval -1 to 1 where values close to 1 indicate a correct clustering: homogenous clusters and well divided. Values close to cero indicates overlapping clusters, and values close to -1 indicate heterogenous clusters and a wrong segmentation. The Calinski-Harabasz index (CH) or Variance Ratio Criterion, developed by Caliński (1974), relates the overall between-cluster variance (i.e the variance between all cluster centroids and the dataset grand centroid) and the overall within-cluster variance (the inertia defined in Equation 6):

$$CH \;=\; \frac{SS_B}{SS_W} \times \frac{n - k}{k - 1} \qquad \text{(10)}$$

---

7   External validation compares the resulting labels form a clustering algorithm with "theoretical" or "true" labels that represent the correct cluster classification.

A high CH indicates that the clusters are relatively well spread out and not too close to each other.

Lastly, the Davies-Bouldin index (DB), proposed by Davies and Bouldin (1979) measures the average degree of similarity between clusters by comparing the distance between clusters and the size of them. More formally it is defined as follows. To measure the similarity between cluster and cluster , we define the degree of similarity as follows:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad (11)$$

Where is the average distance between each object in cluster and its centroid and is the distance between the centroids of both clusters. The DB index is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \qquad (12)$$

Lower and close to cero values of the DB index, indicate a better partition of the groups relative to the entire dataset.

Comparing the informality definition with clustering

Given our objective of checking the relation between the traditional informality definitions with the results of the clustering, we use a simple test for contingency tables. For this sort of external validation test, the null and alternative hypothesis are as follows:

H_0: (Null Hypothesis): There is no relationship between variable 1 and variable 2

H_1: (Alternative Hypothesis): There is a relationship between variable 1 and variable 2

The test statistic relates observed frequencies with the expected frequencies in the absence of a relationship between the categorical variablesin each one of the k cells of the contingency table of r rows and c columns:

$$\chi_d^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad (13)$$

Where the statistic χ_d^2 follows a chi-squared with $d = (r-1)(c-1)$ degrees of freedom.

Empirical results

In this section, we start by presenting the results of the optimal selection of the number of clusters for each algorithm, we move to the internal validation of the three algorithms, later we describe the intra-cluster homogeneity by a series of descriptives of each cluster. Lastly, we compare the labels generated by our preferred clustering algorithm and those resulting from the traditional definitions of informality.

Optimal number of clusters

We now turn to the FAMD-K-Means, K-Modes, and K-Prototypes algorithms optimal number of clusters. As discussed before, for choosing the number of clusters we follow the Elbow Method. The Figure 1 presents the cost function for the three algorithms and different numbers of clusters. The second y-axis corresponds to the inertia measure for the K-Means algorithm. We observe the inflection point and thus the optimal number of clusters is between three and four: for the FAMD-K-Means and the K-Prototypes algorithm, the inflection point is at three clusters and for the K-Modes, the inflection point is at four clusters.

Figure 1. Elbow Method
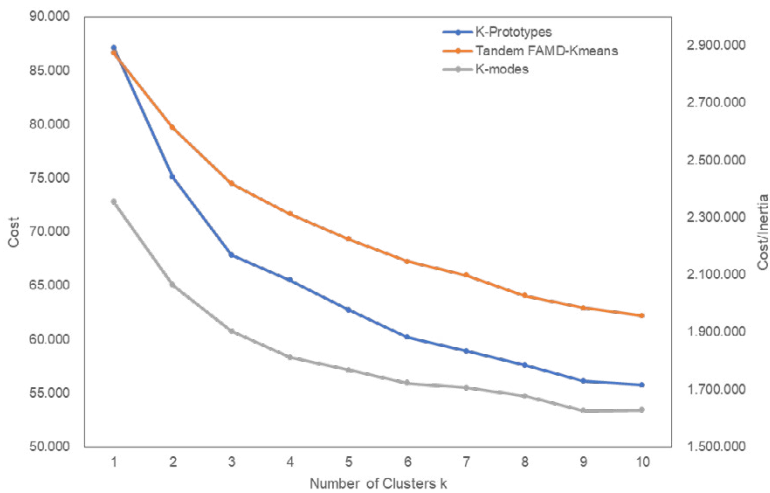
Source: Own elaboration.

These results are similar to the ones found by López-Roldan and Fachelli (2021) or Howell (2011) for other developing economies.

Internal clustering validation

To determine the best algorithm for clustering the Colombian workers data, Table 2 presents the three abovementioned internal validity metrics. Considering that the FAMD-K-Means has the higher Silhouette and Calinski indexes, and the lowest Davies-Bouldin index, it seems that is the best algorithm for the task at hand. It is followed by the K-Modes algorithm with the K-Prototypes in the last place.

Table 2. Internal validation measures for the clustering algorithms

Algorithm	Clusters	Indices		
		S	CH	DB
K-Prototypes	3	0.06	22480.0	4.16
K-Modes	3	0.07	25.88.7	3.65
Tandem: FAMD-K-Means	3	0.13	36591.6	2.29

Source: Own elaboration.

Analysing segments for the Tandem Clustering: FAMD-K-means

Table 3 presents some descriptive statistics for the workers in each of the FAMD-K-means resulting clusters. From the total, the first cluster represents 53.2% of total workers. The second cluster 37.3% and the last cluster 9.5%. The results in Table 3 are shown as a proportion of each category within the cluster in the case of categorical variables with the exception of age and earnings variables for which some central tendency and dispersion statistics are presented within the cluster.

Table 3. Descriptive statistics for the three clusters of workers in the Colombian
labour market using the Tandem Clustering: FAMD-K-Means algorithm

Variable	Category /Statistical	Segments		
		1	2	3
Total observations		**168512**	**118049**	**30001**
(%)		53,2	37,3	9,5
Gender	Male	54.0	59.0	52.0
	Female	46.0	41.0	48.0
Age	Mean	31	54	45
	Std	8	9	11
	Min	10	30	18
	25%	25	47	36
	50%	30	53	44
	75%	36	60	53
	Max	55	98	98

Variable	Category /Statistical	Segments		
		1	2	3
Ethnicity	No ethnicity	90.0	89.0	90.0
	Black, mulato, afrocolombian	8.0	8.0	9.0
	Indigenous	2.0	3.0	1.0
	Palenquero from San Basilio	0.0	0.0	0.0
	Raizal from San Andres, Providencia, Santa Catalina	0.0	0.0	0.0
	Romani	0.0	0.0	0.0
Nationality	Foreigner	7.0	1.0	1.0
	Native	93.0	99.0	99.0
Education level	Without Educ.	1.0	7.0	0.0
	Primary	6.0	42.0	1.0
	Secondary	11.0	19.0	1.0
	Middle school	43.0	22.0	7.0
	Technical of technological ed	28.0	6.0	11.0
	Undergraduate	11.0	4.0	39.0
	Postgraduate	1.0	1.0	42.0
Graduated	Not	17.0	67.0	1.0
	Yes	83.0	33.0	99.0
Working and studying	Not	90.0	100.0	93.0
	Yes	10.0	0.0	7.0
Type of work	Domestic worker	3.0	6.0	0.0
	Day laborer or peon	1.0	3.0	0.0
	Worker or employee of a private company	56.0	21.0	27.0
	Government worker or employee.	1.0	1.0	47.0
	Other work	0.0	0.0	0.0
	Employer	2.0	5.0	6.0
	Self-employed	38.0	65.0	20.0
Contract type and duration	fixed ≤ 6 months	7.0	1.0	6.0
	fixed > 6 months	11.0	3.0	14.0
	Indefinite	26.0	10.0	66.0
	Other	56.0	86.0	13.0
Tenure at the current firm	< 1 year	47.0	21.0	14.0
	2 to 3 years	25.0	14.0	15.0
	4 to 10 years	23.0	26.0	28.0
	11 to 20 years	5.0	20.0	23.0
	> 20 years	0.0	18.0	20.0
Industry	Agriculture and Fishing	4.0	13.0	1.0
	Mining, Manufacturing and Utilities	14.0	13.0	7.0
	Construction	7.0	9.0	3.0
	Wholesale and retail trade	23.0	25.0	5.0
	Hotels and restaurants	9.0	7.0	1.0
	Transport and communication	10.0	10.0	2.0

Variable	Category /Statistical	Segments		
		1	2	3
Industry	Financial intermediation	2.0	0.0	3.0
	Real estate and business activities	8.0	7.0	8.0
	Public administration and defence	2.0	1.0	27.0
	Education	3.0	1.0	32.0
	Health and social work	14.0	8.0	11.0
	Other	3.0	6.0	0.0
Occupation	Senior officials and managers	8.0	14.0	14.0
	Professionals	7.0	2.0	61.0
	Technicians and associate professionals	9.0	2.0	5.0
	Clerks	10.0	3.0	8.0
	Service and sales workers	31.0	24.0	9.0
	Skilled agricultural	4.0	13.0	0.0
	Craft and trades workers	14.0	21.0	1.0
	Plant and machine operators	13.0	13.0	1.0
	Elementary occupations	5.0	9.0	0.0
Working time	<30_part-time	12.0	18.0	4.0
	30-50_full-time	66.0	57.0	84.0
	>50_extra-time	22.0	25.0	12.0
Labour income (mill COP monthly)	Mean	0.85	0.68	3.77
	Std	0.5	0.5	3.3
	Min	0.0	0.0	0.0
	25%	0.5	0.3	2.2
	50%	0.8	0.6	3.0
	75%	1.0	0.9	4.0
	Max	4.5	4.4	100.0

Source: Own elaboration.

The profiles for the three clusters are as follows:

Cluster one (53.2%): *Young workers with secondary or technical education, mostly private workers and self-employed, without a contract, and earnings close to the minimum wage.* This segment is predominantly composed by males (54%) with a median age of 30 years (IQR 25-36 years). 42% of them with secondary education and 28% with technical or technological education, mostly graduated (83%).

They usually work as employees of private firms (56%) but another 38% of them are self-employed. They usually do not have a labour contract (56%) with a reported experience at the current job of less than a year (47%) they typically work as a salesperson (31%) in the commerce industry (23%). Their average earnings are around COP $850 thousand in 2019. Slightly above the monthly minimum wage of that year COP $828 thousand.

Cluster two (37.3%): *Adult workers with basic education, mostly self-employed, with earnings below a monthly minimum wage.* The segment is predominantly composed by males (59%) with a median age of 53 years (IQR 47-60 years), mainly with primary education (42%) and curiously all of them are not currently studying.

They usually work as self-employed (65%), without a labour contract (86%) and report working by themselves (61%) in commerce (25%) and agriculture and fishing (13%) with monthly earnings around COP $680 thousand below the monthly minimum wage of that year COP $828.

Cluster three (9.5%): *Adult workers, with higher education, with indefinite labour contract and high earnings.* This group is almost equally divided by gender. A median age of 44 years (IQR 36-53 years). Highest educational achievement is undergraduate degree (39%) postgraduate degree (42%) and almost all of them got a degree (99%), 7% of them study and work at the same time.

In terms of labour demand, this group is comprised of private workers (27%) or government workers (47%), 6% of them are employers (the highest share among the clusters) and only 20% are self-employed. Most of them have an indefinite labour contract (66%) with tenure at the current job between 3 and 20 years (51%). They typically work in the education (32%) public administration and defence (27%) industries. Their occupation is mostly professionals (61%), 84% working full time and they earn way above the monthly minimum wage at COP $3.77 million.

These results are similar to the ones found by Howell (2011), where there were clear differences in earnings and education between the three segments found for the capital of Xianjing in China.

Contrasting the clustering approach with the informality definition

In this section, we compare the labels obtained with the clustering algorithm and those using two traditional definitions of labour informality in Colombia. The first one is the legalistic view; it verifies if the worker is currently making pension contributions (formal) or not (informal). The second definition is the productivity view, a worker in a firm of five workers or more, or a professional independent worker are defined as formal, any other worker is defined as informal. The proposed contingency Tables 4 and 5 are presented below.

Table 4. Clustering and the legalistic view of labour informality

1		Tandem: FAMD-K-Means			Total
		2	3		
Contributing to pension insurance	Formal	67387	48316	12247	127.950
	Informal	99817	70209	18586	188.612
Total		167.204	118525	30833	316562

Source: Own elaboration.

Table 5. Clustering and the productivity view of labour informality

1		Tandem: FAMD-K-Means			Total
		2	3		
Firm's size	Formal	71118	51067	13097	135.282
	Informal	96086	67458	17736	181.280
Total		167.204	118525	30833	316562

Source: Own elaboration.

From the results presented before, it seems that the third cluster contains mostly formal workers, with high earnings, high education, and indefinite labour contracts. The first cluster, comprising workers with vocational training, working as employees seemed to have some formal workers while the second cluster seemed to have mostly informal workers, with low earnings and mainly self-employed. However, Tables 4 and 5 indicate that the three clusters have formal and informal workers alike. This is especially true if we consider that the third cluster has a higher proportion of informal workers than formal, despite the descriptive statistics for this cluster presented in the previous section.

Considering a formal test of correlation between the traditional definition of informality and our cluster, we find a clear rejection of the independence between the two categorical variables. The Chi-squared test presented at section 3.2.6 has two degrees of freedom in both cases. In the case of the legalistic view of informality, the value of the statistic is around 13, which implies that we reject the null hypothesis of independence between the variables even at the 1% significance level. For the productivity definition of informality, the corresponding statistic is 9.5 which also implies that we reject the null hypothesis of independence at the 1% significance level. These results imply that the segments resulting from the Tandem: FAMD-K-Means are correlated to the ones resulting from traditional definitions of labour informality.

Conclusions

In this article, we propose to use an unsupervised machine-learning algorithm to test whether the Colombian labour market is segmented or must be considered as one for policymaking. To introduce the top-down clustering methodology to a non-data scientist audience, especially of labour economist, we describe in detail three algorithms with their pre-processing requirements; we compared them using three internal validation indicators. We found that Tandem: FAMD-K-Means is the best algorithm giving the higher intragroup homogeneity and extra group heterogeneity relative to the other two algorithms.

In an external validation, we also compared the partition resulting from our algorithm with the two most employed definitions of informality: The legalistic view that assumes the worker is

formal if he is contributing to social security, in our context pension contribution. The productivity view: large firms, measured in workers numbers are formal. Despite in simple contingency tables we find no clear pattern between the two approaches to segmentation, we reject the null hypothesis of independence of our clustering with each one of the definitions.

The results indicates that the labour market is segmented and in labour policy terms one size does not fit all. Our results are in line with other employment quality research: the group of quality jobs in Colombia (in our case the third cluster) is much reduced and earnings differentials are crucial classifying jobs (Farné et al., 2013). There is also a large group of workers (first cluster) with earnings around the minimum wage, mostly composed of young adults, most of them without a university degree but with vocational education, half of these workers are found to be contributing to social security or in large firms, this implies that there is some formalisation potential among this group. Lastly, there is another large group of workers (second cluster) mostly older without tertiary education, with earnings below the minimum wage, this last group has lesser formalisation potential than the previous group.

The comparison between our clusters and the traditional informality definitions could imply that the simple indicators are just one dimension to take into account in analysing segmentation in developing economies, and that the problem is more intricated that what is typically depicted in labour economics models.

We consider this alternative measurement of segmentation an important endeavour, the clustering approach could improve public policy by allowing policy makers to devise strategies focused on specific and more homogenous groups. From a methodological point of view, future research could consider checking the appropriateness of bottom-up algorithms; making a comparison in terms of the internal validation indicators of clustering and traditional definitions of informality and lastly, exploring alternative methods to determine the optimal number of clusters different to the Elbow method.

References

[1]   Aaberge, R., Dagsvik, J. K., & Strøm, S. (1995). Labor Supply Responses and Welfare Effects of tax Reforms. *The Scandinavian Journal of Economics*, 635-659. https://doi.org/10.2307/3440547

[2]   Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering: Algorithms and Applications*. CRC Press.

[3]   Ahmad, A., & Hashmi, S. (2016). K-Harmonic Means Type Clustering Algorithm for Mixed Datasets. *Applied Soft Computing, 48*, 39-49. https://doi.org/10.1016/j.asoc.2016.06.019

[4]   Akay, Ö., & Yüksel, G. (2018). Clustering the Mixed Panel Dataset Using Gower's Distance and K-Prototypes Algorithms. *Communications in Statistics-Simulation and Computation, 47*(10), 3031-3041. https://doi.org/10.1080/03610918.2017.1367806

[5]   Anderson, K. H., Butler, J. S., & Sloan, F. A. (1987). Labor Market Segmentation: A Cluster Analysis of Job Groupings and Barriers to Entry. Southern Economic *Journal, 53* (3), 571-590. https://doi.org/10.2307/1058755

[6]   Apitzsch, C., & Ryeng, J. (2020). *Cluster Analysis of Mixed Data Types in Credit Risk: A study of clustering algorithms to detect customer segments* [M.Sc. Thesis, Umea University]. https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-172594

[7]   Aschenbruck, R., & Szepannek, G. (2020). Cluster Validation for Mixed-Type Data. *Archives of Data Science, Series A, 6*(1), 02. https://publikationen.bibliothek.kit.edu/1000120412/79692380

[8]   Bock, H. H. (1987). *On the Interface between Cluster Analysis, Principal Component Analysis, and Multidimensional Scaling.* In Multivariate Statistical Modeling and Data Analysis: Proceedings of the Advanced Symposium on Multivariate Modeling and Data Analysis May 15–16, 1986 (pp. 17-34). Springer Netherlands.

[9]   Boston, T. D. (1990). Segmented Labor Markets: New Evidence from A Study of Four Race-Gender Groups. *ILR Review, 44*(1), 99-115. https://doi.org/10.2307/2523432

[10]   Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27. https://doi.org/10.1080/03610927408827101

[11]   Cao, F., Liang, J., & Bai, L. (2009). A New Initialization Method for Categorical Data Clustering. *Expert Systems with Applications, 36*(7), 10223-10228. https://doi.org/10.1016/j.eswa.2009.01.060

[12]   Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

[13]   Departamento Nacional de Estadística. (2019). *Gran Encuesta Integrada de Hogares (GEIH)* [data base] https://microdatos.dane.gov.co/index.php/catalog/599

[14]   de Soto, H. (1989). *The Other Path: The Invisible Revolution in the Third World.* Basic Books.

[15]   Farné, S. (1990). Segmentación laboral y ciclo económico. Una tentativa de estimación para el caso colombiano. *Desarrollo y Sociedad*, 1(25), 89-122. https://doi.org/10.13043/dys.25.3

[16]   Farne, S., Rodríguez D., & Carvajal, Y. (2013). *La calidad del empleo en 23 ciudades colombianas* [bulletin]. Boletín 14 Observatorio Laboral. Universidad Externado de Colombia. https://ideas.repec.org/p/col/000194/015977.html

[17]   Gittleman, M. B., & Howell, D. R. (1995). Changes in the Structure and Quality of Jobs in The United States: Effects by Race and Gender, 1973–1990. *ILR Review, 48*(3), 420-440. https://doi.org/10.1177/001979399504800303

[18]   Gong, X., van Soest, A., & Villagomez, E. (2004). Mobility in the Urban Labor Market: a Panel Data Analysis for Mexico. *Economic Development and Cultural Change, 53*(1), 1-36. https://doi.org/10.1086/423251

[19]   Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 857-871. https://doi.org/10.2307/2528823

[20]   Grané, A., & Sow-Barry, A. A. (2021). Visualizing Profiles of Large Datasets of Weighted and Mixed *Data. Mathematics,* 9(8), 891. https://doi.org/10.3390/math9080891

[21]   Harris, J. R., & Todaro, M. P. (1970). Migration, Unemployment and Development: A Two-Sector Analysis. *The American Economic Review, 60*(1), 126-142. https://www.jstor.org/stable/1807860

[22]   Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction,* 2nd edition. Springer.

[23]   Henley, A., Arabsheibani, G. R., & Carneiro, F. G. (2009). On Defining and Measuring the Informal Sector: *Evidence from Brazil. World development, 37*(5), 992-1003. https://doi.org/10.1016/j.worlddev.2008.09.011

[24]   Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of Cluster Analysis.* CRC press.

[25]   Howell, A. (2011). Labor Market Segmentation in Urumqi, Xinjiang: Exposing Labor Market Segments and Testing the Relationship between Migration and Segmentation. *Growth and Change, 42*(2), 200–226. https://doi.org/10.1111/j.1468-2257.2011.00550.x

[26]   Hussmanns, R. (2004). Measuring the Informal Economy: From Employment in the Informal Sector to Informal *Employment* [Bureau of Statistics, International Labour Office Working Paper No. 53]. https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@integration/documents/publication/wcms_079142.pdf

[27]   Huang, Z. (1997a). Clustering Large Data Sets with Mixed Numeric and Categorical Values. Proceedings of the First Pacific-Asia Conference. Singapore, World Scientific (pp. 21-34). https://www.scirp.org/reference/referencespapers?referenceid=1497127

[28]   Huang, Z. (1997b). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Dmkd, 3*(8), 34-39. https://www.semanticscholar.org/paper/A-Fast-Clustering-Algorithm-to-Cluster-Very-Large-Huang/a17726899429fc6c8a92779556efe43d9472b7b5

[29]   Huang, Z. (1998). Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery, 2*(3), 283-304. https://doi.org/10.1023/A:1009769707641

[30]   International Labour Office (ILO). (1972). *Employment, Incomes and Equality: A Strategy for Increasing Productive Employment in Kenya.* ILO.

[31]   Koren, O., Hallin, C. A., Perel, N., & Bendet, D. (2019). Enhancement of the K-Means Algorithm for Mixed Data in Big Data Platforms. In K. Arai, S. Kapoor & R. Bathia (eds.), *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys),* Volume 1 (pp. 1025-1040). Springer International Publishing.

[32]   López-Roldán, P., & Fachelli, S. (2021). Measuring Labour Market Segmentation for a Comparative Analysis Among Countries. *Social Indicators Research, 154*(3), 857-892. https://doi.org/10.1007/s11205-020-02550-1

[33]   Magnac, T. (1991). Segmented or Competitive Labor Markets. *Econometrica, 59*(1), 165-187. https://doi.org/10.2307/2938245

[34]   Maloney, W. F. (1999). Does Informality Imply Segmentation in Urban Labor Markets? Evidence from Sectoral Transitions in Mexico. *The World Bank Economic Review, 13*(2), 275-302. https://www.jstor.org/stable/3990099

[35] Martin, C., & Okolo, M. (2022). *Heterogeneity in the UK Labour Market: Using Machine-learning To Test Macroeconomic Models*. (Bath Economics Research Papers; No. 93/22). Department of Economics, University of Bath. https://researchportal.bath.ac.uk/en/publications/heterogeneity-in-the-uk-labour-market-using-machine-learning-to-t

[36] Modha, D. S., & Spangler, W. S. (2003). Feature Weighting in K-means Clustering. *Machine Learning, 52*, 217-237. https://doi.org/10.1023/A:1024016609528

[37] Neffa, J. C. (2008). Las teorías de la segmentación de los mercados de trabajo. In F. Eymard-Duvernay & Neffa (eds.), *Teorías económicas sobre el mercado de trabajo* (pp. 139-206). Fondo de Cultura Económica.

[38] Neffa, J. C. (2023). Teorías de la segmentación del mercado de trabajo. *RBEST: Revista Brasileira de Economia Social e do Trabalho, 5*(00), e023012. https://doi.org/10.20396/rbest.v5i00.18343

[39] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-Learn: Machine-learning in Python. *Journal of Machine-Learning Research, 12*, 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

[40] Perry, G. E., Arias, O., Fajnzylber, P., Maloney, W. F., Mason, A., & Saavedra-Chanduvi, J. (2007). *Informality: Exit and Exclusion*. World Bank.

[41] Pradhan, M., & van Soest, A. (1995). Formal and Informal Sector Employment in Urban Areas of Bolivia. *Labour Economics, 2*(3), 275-297. https://doi.org/10.1016/0927-5371(95)80032-S

[42] Pradhan, M., & van Soest, A. (1997). Household Labor Supply in Urban Areas of Bolivia. *Review of Economics and Statistics, 79*(2), 300-310. https://doi.org/10.1162/003465397556656

[43] Rodríguez, D. (2021) Microsimulation Analysis of Informal Labour Markets in Developing Countries (PhD thesis, University of Essex). http://repository.essex.ac.uk/id/eprint/29928

[44] Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to The Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[45] Sousa-Poza, Alfonso. (2004). Is the Swiss Labor Market Segmented? An Analysis Using Alternative Approaches. *Labour, 18*(1), 131–161. https://doi.org/10.1111/j.1121-7081.2004.00261.x

[46] van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019). Distance-Based Clustering of Mixed Data. *Wiley Interdisciplinary Reviews: Computational Statistics, 11*(3), e1456. https://doi.org/10.1002/wics.1456