Indicator for the Regional Labor Market Using Machine Learning Techniques: Application to Colombian Cities[*]

Received: April 25, 2024 – Accepted: March 3, 2025

Doi: https://doi.org/10.12804/revistas.urosario.edu.co/economia/a.14392

Pavel Vidal[†]
Lya Paola Sierra-Suárez[‡]
Julieth Cerón[§]

Abstract

This article proposes a methodology to estimate a labor market indicator that combines economic, social, inequality, and expectation variables. Machine Learning techniques are used to select the most relevant variables. The indicator captures the traditional evolution of the employment and unemployment rates and incorporates information on gender, age, informality, productive sectors, and Google Trends data. This approach allows for a more comprehensive understanding of the labor market situation, better visibility of

[*] We are grateful for the collaboration of Miguel Velasquez and Christian Valor in the development of this research.

[†] Department of Economics at the Pontificia Universidad Javeriana, Cali, Colombia. Correo electrónico: pavel@javerianacali.edu.co. ORCID: https://orcid.org/0000-0001-8278-3122

[‡] Department of Economics at the Pontificia Universidad Javeriana, Cali, Colombia. Correo electrónico: lyap@javerianacali.edu.co. ORCID: https://orcid.org/0000-0002-8909-8977

[§] Department of Economics at the Pontificia Universidad Javeriana, Cali, Colombia. Correo electrónico: stefens07@ javerianacali.edu.co. ORCID: https://orcid.org/0000-0002-5365-2534

Para citar este artículo: Vidal, P., Sierra-Suárez, L. P., & Cerón, J. (2025). Indicator for the Regional Labor Market Using Machine Learning Techniques: Application to Colombian Cities. *Revista de Economía del Rosario*, *27*(1), 1-31. https://doi.org/10.12804/revistas.urosario.edu.co/economia/a.14392

ISSN 0123-5362 - ISSNE 2145-454x

regional differences, and analysis of the heterogeneous impact of the pandemic and subsequent recovery. The methodology is exemplified in the Colombian cities of Cali, Medellín, Bogotá D.C., and Popayán.

Keywords: labor market indicator; machine learning; Lasso; backward stepwise selection method; principal components; Google Trends.
Classification JEL: J21, C45, C38, C55, C88

Indicador para el Mercado Laboral Regional Usando Técnicas de Aprendizaje Automático: Aplicación a Ciudades Colombianas

Resumen

Este artículo propone una metodología para estimar un indicador del mercado laboral que combina variables económicas, sociales, de desigualdad y de expectativas. Se emplean técnicas de aprendizaje automático para seleccionar las variables más relevantes. El indicador captura la evolución de las tasas de empleo y desempleo, e incorpora información sobre género, edad, informalidad, sectores productivos y datos de Google Trends. Este enfoque permite una comprensión más integral de la situación del mercado laboral, una mejor visibilidad de las diferencias regionales, así como la heterogeneidad del impacto de la pandemia y la posterior recuperación. La metodología se ejemplifica en las ciudades colombianas de Cali, Medellín, Bogotá D.C. y Popayán.

Palabras clave: indicador del mercado laboral; aprendizaje automático; Lasso; método Backward; componentes principales; Google Trends.

Indicador do mercado de trabalho regional usando técnicas de aprendizado de máquina: aplicação a cidades colombianas

Resumo

Este artigo propõe uma metodologia para estimar um indicador do mercado de trabalho que combina variáveis econômicas, sociais, de desigualdade e de expectativas. Técnicas de aprendizado de máquina são empregadas para selecionar as variáveis mais relevantes. O indicador capta a evolução das taxas de emprego e desemprego, incorporando informações sobre gênero, idade, informalidade, setores produtivos e dados do Google Trends. Essa abordagem permite uma compreensão mais ampla da dinâmica do mercado de trabalho, oferecendo maior visibilidade das diferenças regionais, bem como da heterogeneidade dos impactos da pandemia e da recuperação subsequente. A metodologia é aplicada às cidades colombianas de Cali, Medellín, Bogotá D.C. e Popayán.

Palavras-chave: indicador do mercado de trabalho; aprendizado de máquina; Lasso; método Backward; componentes principais; Google Trends.

Introduction

The labor market is subject to shocks of different natures and signs and is influenced by economic, social, and demographic dynamics. This often leads to the various traditional employment-related indicators moving in contradictory directions and offering divergent signals.

It is complex to have a single and structured reading of labor market conditions in the short term, considering the diverse information and trends offered, for example, by the unemployment rate, the employment rate, the overall participation rate, the number of people employed in different industries, wages, and other indicators by gender, age, or productive sectors. The employment and unemployment rates are expected to almost always move in opposite directions. Still, this rule is violated frequently, making it difficult to interpret the underlying trends of the labor market and the economic cycle, the analysis of the effect of public policies, and private sector decisions.

Efforts to build synthetic indicators of the labor market at the regional or city level within countries are not frequent. An indicator summarizing the aggregated evolution (incorporating economic, social, inequality, and expectation variables) allows for a standardized and comprehensive comparison between regional labor markets.

After the pandemic, cities within Colombia quickly resumed a gradual recovery of the leading labor market indicators. However, the recovery, like the fall during the border closure and physical distancing measures, was heterogeneous. Different authors have studied evidence showing more significant vulnerabilities in certain social groups and economic sectors during the pandemic, for example, among young people, women, workers in the trade, hospitality, and tourism sector, in informal employment, and micro and small enterprises (Adams-Prassl et al., 2021; Alon et al., 2021; Bonilla & Gaviria, 2020; Cajner et al., 2021).

Velasco (2021) analyzes that, in Latin America during the pandemic, the reduction in labor participation was much more pronounced among women, and the destruction of informal jobs was more significant than the contraction of formal employment, unlike previous crises. The author also examines how informal employment absorbed part of the loss of salaried and formal jobs. In Colombia, Morales et al. (2022) found that salaried workers were the most affected by these mobility restrictions, especially in activities where teleworking was impossible. They also conclude that industries with many

small companies and a high proportion of minimum wage employees were more affected during the pandemic.

Such heterogeneities are not new, although some seemed to widen during the pandemic and subsequent economic recovery. At the city level, labor market divergences have been visible, responding to particularities in the productive structure and employment and the policies applied by local authorities.

The heterogeneous evolution of the regional labor market can be broadly appreciated in traditional indicators such as the unemployment rate, participation in the labor force, and employment, among others reported by statistical offices. However, interregional comparisons can lead to different conclusions depending on the indicator considered, as not all show the same trends or gaps when comparing cities.

This article proposes a methodology combining different variables of the regional labor market to develop a monthly indicator that synthesizes the aggregated or average situation, considering traditional employment metrics and variables that consider social and inequality aspects. The indicator captures the evolution of the employment and unemployment rates and contains information on gender, age, informality, productive sectors, and expectations. In this way, a more comprehensive understanding of the labor market situation can be achieved, better highlighting regional differences and analyzing the heterogeneous impact of the pandemic and subsequent recovery.

Regarding the methodology, the indicator has two notable aspects. First, Google searches related to the labor market are used, which allows for considering public perceptions and expectations regarding employment and incorporating the indicator within the emerging data analysis trend by using information from social networks and the Internet. Second, Machine Learning (ML) methods are utilized; in particular, Lasso regression and "Backward Stepwise" are used to select the most relevant variables in the indicator estimation. Machine Learning techniques and the Dynamic Factor Model are applied to extract the comovement among the selected variables using Principal Components.

The methodology is exemplified in the Colombian cities of Cali, Medellín, Bogotá D.C., and Popayán. To date, the authors are not aware of any similar exercises developed to assess the labor market at the city level using these methodologies. At a country level, Orozco-Castañeda et al. (2024) developed a labor market indicator using a methodology similar to previous models. However, they focused on forecasting labor market outcomes through machine learning techniques. By incorporating diverse data

sources, including Google search data, their approach aimed to enhance the accuracy of Colombia's employment and unemployment rate predictions, providing more precise labor market assessments during the unprecedented times of Covid-19. Furthermore, Vidal et al. (2017) have utilized the factor model to estimate the Monthly Economic Activity Indicator (IMAE), which has been applied across various regions, including Valle del Cauca, Cauca, Antioquia, Cali, Caribe, and Nororiente. For applications of the IMAE in Colombia, refer to Sierra-Suarez et al. (2022) and Orozco-Gallo et al. (2021). For a more detailed explanation of the methodology used to create economic activity indexes using factor models, see Sierra-Suárez et al. (2017).

The selection of Cali, Medellín, Bogotá D.C., and Popayán for this study is strategic, as economic activity indicators (IMAES) are already established for these departments. This existing data enables seamless integration and enhances complementarity with ongoing studies examining the relationship between labor indicators and economic activity at a regional level. In addition to the research and analysis phase, the construction of the indicators has the potential to be updated monthly as a tool for systematically monitoring the labor market and analyzing the economic situation in the cities.

The rest of this document is organized as follows: Section 2 presents a brief review of the literature on the importance of new sources and data management techniques for the labor market; Section 3 describes the evolution of the labor market in the cities under analysis; sections 4 and 5 describe the methodology and the data, respectively; Section 6 presents the results; and Section 7 contains the conclusions.

Literature Review: New Information Sources for the Labor Market and Machine Learning Techniques

Various measures can explain the labor market's performance, including the unemployment rate, the employment rate, the labor force participation rate, different metrics on labor underutilization, long-term unemployment, unemployment flows by social groups, the average weekly hours worked, and wage growth (Zmitrowicz & Khan, 2014).

The unemployment rate is usually considered the metric that best describes the labor market conditions (Dvorkin, 2015; Zmitrowicz & Khan, 2014). However, it is recognized that the unemployment rate is just one aspect of the labor market. Many other characteristics of labor market conditions are no less important. A more extensive list of crucial indicators for the labor market can be found, for example, in the works of Ramos-Veloza et al. (2019) and Hakkio and Willis (2013). The challenge when moving towards

a more inclusive information approach is the complexity of combining and simultaneously interpreting the disparate indicators that approximate the labor market conditions from different angles, scales, and frequencies.

The reports from the National Administrative Department of Statistics (DANE) of Colombia generally highlight three indicators as the core of labor market analysis, measured from the Great Integrated Household Survey (GEIH): the unemployment rate (proportion of the labor force that is currently not occupied), the labor participation rate (percentage of the working-age population that is employed or actively seeking employment), and the employment rate (number of people employed compared to the total working-age population). All these are standardized metrics used internationally. However, DANE systematically calculates and publishes a broader range of labor market indicators that analysts, media, and the public increasingly consider.

The construction of indicators synthesizing the characteristics and trends of an economy, a market, or an industry is not new. Different methods have been refined and explored for these purposes. One of the most used methods to build indices or latent factors is the Dynamic Factor Model (DFM), which can be estimated using principal components or the Kalman filter (Chung et al., 2015; Sierra-Suárez et al., 2017; Swanson & Xiong, 2018; Vidal et al., 2017).

The construction of synthetic indicators has benefited from the expansion of Web 2.0, which represented a significant change in how the Internet is used, moving from a platform of passive consumption to one of user participation. In this new paradigm, users create and share content, interact, and collaborate through dynamic web platforms and applications. This has resulted in an explosion of user-generated data, which are collected, analyzed, and used for various purposes. This phenomenon has led to the era of "Big Data." New information sources sometimes provide data volumes that are too complex for traditional processing methods. Big Data is essential for many modern applications, such as Machine Learning (ML), predictive analytics, marketing personalization, and data-driven decision-making.

Due to data capture, storage, and computational power advances in the last decade, ML tools have become crucial for managing large data sets (Baldacci et al., 2016; Medeiros et al., 2019). One of the virtues of ML methods is that, when working with large data sets and potential variables, they allow for the reduction of the database's dimension following specific statistical selection criteria. Other important properties are their effectiveness in modeling complex relationships and their application across multiple fields (Varian, 2014).

In recent years, ML techniques such as "Backward Stepwise" selection and Lasso regression ("Least Absolute Shrinkage and Selection Operator") have become popular for selecting a relevant subset of explanatory variables and reducing the complexity of the model. Choi and Shin (2019) demonstrated that data differentiation and dimensionality reduction with the Lasso method improved prediction accuracy and computational efficiency. Similar results have been valued with the "Backward Stepwise" technique.

Within macroeconomic forecasting, selection methods are widely used to decrease the number of predictors utilized. Chen and Wu (2023) adopted an innovative approach by combining the Lasso method with the ARIMA model to forecast Chongqing's GDP. They showed that Lasso selected key economic features and indicators linked to Chongqing's GDP, thereby improving the accuracy of short-term forecasts. Jokubaitis et al. (2021) integrated the Lasso technique with principal components to enhance the prediction of GDP components in the United States and the European Union. Their research showed that the Lasso selection method outperformed conventional approaches based on principal components, significantly identifying more effective predictor sets.

In their study, Qin et al. (2022) created an algorithm based on supervised dimensionality reduction for macroeconomic predictions. Their conclusion highlighted that the "Backward Stepwise" selection method allowed for choosing the optimal set of financial indicators, one-dimensional, partial, and leading variables of macroeconomic variables, from an initially broad set. This subset proved most effective for short-term macroeconomic forecasts in the United States.

The computational methodology of ML is increasingly applied in economics and finance, prediction exercises, and indicator construction. New approaches seek to compete with traditional models and offer better outcomes. Various authors conclude that ML techniques perform better when there is a larger data set than traditional estimation methods (Baldacci et al., 2016; Mullainathan & Spiess, 2017; Swanson & Xiong, 2018; Varian, 2014).

Another trend in data analysis incorporated in the labor market indicator in this article is information from Google Trends. One of the first studies to test the validity of using Google Trends data to predict the labor market was by Choi and Varian (2012). Caperna et al. (2020) studied the impact of the Covid-19 pandemic in 27 European Union countries using Google Trends data to predict the unemployment rate. Simionescu and Raišienė (2021) used Google Trends data to measure the impact of the pandemic on employment expectations. Jun et al. (2018) examined the growing utility of Google Trends searches in research across numerous areas such as informatics,

communications, medicine, health, business, and economics. Their ten-year review of research using Google Trends finds that the approach has moved from merely describing trends to using these data to build projections.

In Colombia, the work of Cardona and Rojas (2017) must be mentioned, which proposed using Google Trends data to improve short-term predictions of national unemployment rates. The authors selected Google search terms that were most related to unemployment rates and estimated simple linear regression models, ARIMA models, and ARIMA models with exogenous variables. The study found that search volume improved model fitness and that searches for terms like "Job," "Job offers," and "Looking for a job" improved predictions of labor market behavior. The analysis concluded that Google Trends data can be a valuable tool for predicting unemployment rates in Colombia.

In Latin America, the work of Campos and López-Araiza (2020) stands out, as they explored the use of Google Trends data and Machine Learning algorithms to predict the unemployment rate in Mexico. The results showed that the Lasso method outperformed an autoregressive model, even when incorporating Google data, and that the random forests method performed slightly worse than Lasso. The article highlighted the importance of these methods and new data sources for economic research and policy design.

The Colombian Regional Labor Market According to the Unemployment and Employment Rate

Figures 1 and 2 show, for the four analyzed Colombian cities, the monthly frequency trajectory of the two main metrics usually employed in the labor market: employment rate and unemployment rate. These data represent the comparison point against which the signals provided by the labor market indicator estimated with ML and MFD techniques are later contrasted.

In the employment rate series, greater heterogeneity among the cities can be observed until 2014, and a greater convergence between Medellín, Cali, and Popayán since 2015. During the Covid-19 pandemic, the employment rate for the four cities recorded the most significant drop, particularly during the second quarter of 2020, when the most restrictive measures for border closures and mobility control took place. According to this metric, Popayán was the city that registered the lowest employment rate during this period, falling to 35.1%, the lowest in its history. It was followed by Cali (40.1%), Medellín (43.6%), and Bogotá (46.3%). From the third quarter of 2020, a gradual recovery in the regional employment rate is evident (Figure 1).

In 2021 and 2022, the recovery of the employment rate in the four cities is observed, although with different levels and speeds, once again highlighting significant regional heterogeneities. By the end of 2022, Cali, Medellín, and Bogotá stand out for resuming employment rates over 60%. In contrast, Popayán closed 2022 with a lower employment rate (55.8%).

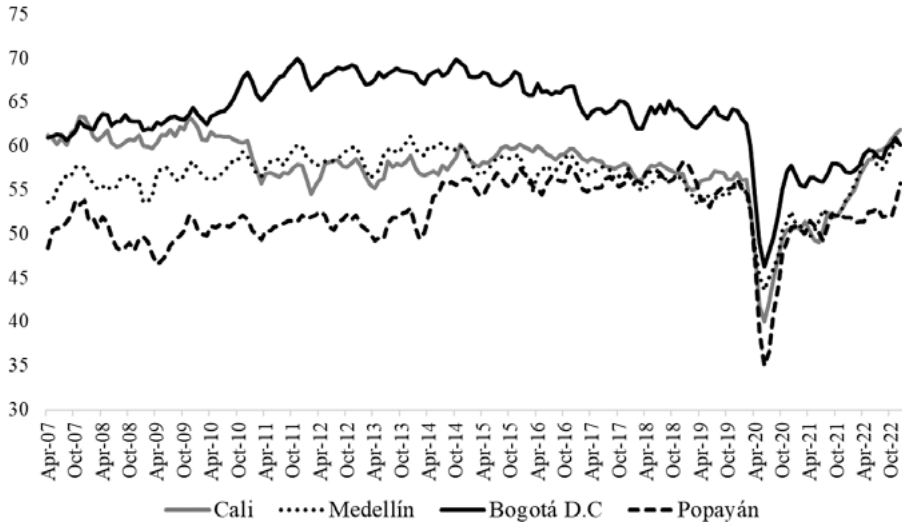

Figure 1. Evolution of the Regional Employment Rate (Monthly Data, 2007-2022)

Source: Authors' elaboration.

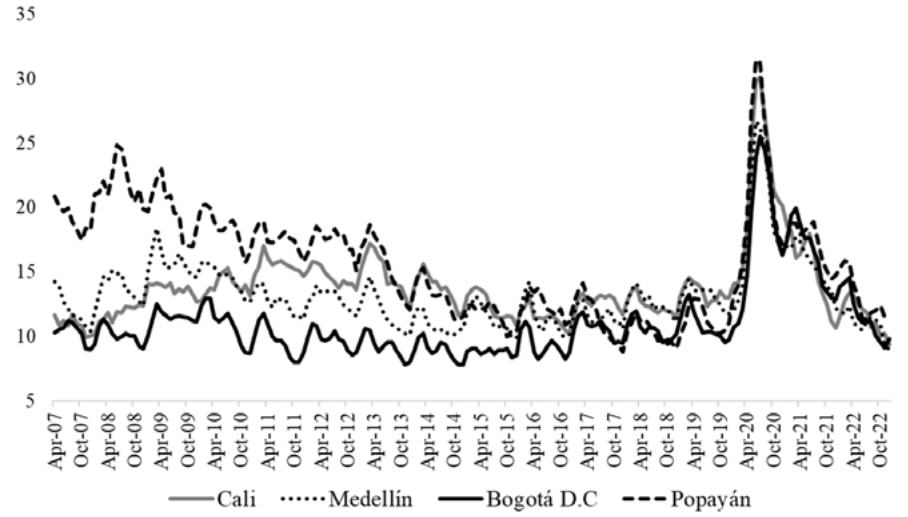

Figure 2. Evolution of the Regional Unemployment Rate (Monthly Data, 2007-2022)

Source: Authors' elaboration.

There are several coinciding signals in the recent historical evolution of these two essential metrics of the Colombian labor market. The regional unemployment rate (Figure 2) also shows greater heterogeneity among the cities until 2014 and greater convergence from 2015. On average, Bogotá has a more robust labor market (a significantly higher employment rate than the other cities examined and a lower unemployment rate), with Popayán at the other extreme. The impact of the pandemic during the same period in the four cities and the subsequent recovery is also visible.

However, there are also some notable differences between these two metrics. For example, the convergence between labor markets since 2015 is much greater when examining the unemployment rate than the employment rate. Before the impact of the pandemic (2019), Cali had the highest unemployment rate, while Medellín had the lowest employment rate. The signals between these metrics during the pandemic and subsequent recovery are also contradictory. While they agree that Bogotá was the city least impacted by the pandemic and Popayán the most affected, according to the unemployment rate, Medellín was the second city affected. In contrast, according to the employment rate, Cali was the second most affected city. In 2021 and 2022, the unemployment rate indicates a much more synchronized recovery than the information offered by the employment rate.

Another event that affected the Colombian labor market in this period was the social protests and road blockades in May and June of 2021. Both the unemployment and employment rates show that Popayán and Cali were the most affected cities, but they disagree on the least affected cities. The unemployment rate suggests that Medellín had the most negligible impact, while the employment rate indicates that it was Bogotá.

Methodology

The regional labor market indicator was created using a two-step methodology: first, variables from the registry built for each city (see Annex A) were selected using ML techniques. Specifically, the "Backward Stepwise" selection and Lasso Regression methods were used. The goal is to identify an optimal subset of variables that later facilitates and improves the estimation process (James et al., 2013). Variables in which both methods coincided were selected.

After selecting the optimal and most relevant number of variables through the two ML methods, the regional labor market indicators were constructed based on the Dynamic Factor Model (MFD).

Backward Stepwise Selection

Backward Stepwise Selection starts with a complete set of potentially predictive variables and iteratively removes the least important ones from a regression model one at a time. The process is based on minimizing the regression model's cost function and selecting the least important variables according to some predefined criterion, which can be the p-value or the adjusted coefficient of determination, also called adjusted R-squared (R^2) (Guyon and Elisseeff, 2003; Hastie et al., 2009; James et al., 2013; Miller, 2002).

The cost function is defined in equation 1 as:

$$Costo = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_{i)}{}^2 \tag{1}$$

where,

n is the number of observations
y_i is the observed value of the response variable
$\hat{y}_i$ is the value predicted by the regression model

Lasso Regression

Lasso Regression is a regression and variable selection technique used to estimate linear models while simultaneously performing automatic variable selection by reducing some coefficients to zero. The method aims to minimize the cost function, which includes a penalty on the sum of the absolute values of the coefficients. This leads to automatic variable selection, as some coefficients are reduced to zero. The penalty value controls the degree of constraint on the magnitude of the coefficients (Tibshirani, 1996).

The goal is to minimize the cost function, which involves finding the optimal values for the coefficients β_j and the optimal value of λ. As the value of λ increases, more coefficients will be reduced to zero, leading to automatic variable selection by identifying those most important for prediction (Hastie, et al., 2009; Tibshirani, 1996).

The cost function is summarized in Equation 2 as:

$$Cost = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_{i)}{}^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{2}$$

where,

n is the number of observations
p is the number of predictor variables
y_i is the observed value of the response variable
$\hat{y}_i$ is the value predicted by the linear model
β_j is the regression coefficient associated with the predictor variable.
λ is the regularization or penalty parameter

Dynamic Factor Model

The Dynamic Factor Model (DFM) assumes that a matrix Y_t of N observed variables can be represented as the sum of two unobservable components that are mutually independent: a common component to all variables F_t, and an idiosyncratic component μ_t which represents the unique dynamics of each series (Sierra-Suárez et al., 2017).

Thus, the equation for the observed series, Y_t can be represented in vector form as:

$$Y_t = P\,F_t + \mu_t \tag{3}$$

where,

Y_t is a vector with t
he observed labor market variables
P is the factor loading matrix (weights)
F_t is the common factor (labor market indicator)
μ_t which represents the unique dynamics of each series.

The Principal Components (PC) method was used to estimate the common factor (Jolliffe, 2002; Jolliffe & Cadima, 2016; Sierra-Suárez et al., 2017). PC has a long history of estimating economic activity indicators (Chung et al., 2015; Hakkio & Willis, 2013; Stock & Watson, 2012).

Data

The starting point for the labor market indicator was the construction of a database as broad as possible, with variables related directly and indirectly to the labor market in each of the four study cities, which were available on a monthly frequency. For Cali, Medellín, and Bogotá, 53 variables were

collected, and for Popayán, 51 variables for 2007-2022. The primary source of the data was DANE, through the GEIH, which collects monthly information on employment, the workforce, and the sociodemographic characteristics of the Colombian population (Annex A).

Information on the main search terms related to the dynamics of the labor market in each city of analysis was extracted from the Google Trends platform. This platform reports the search indices for any word in the browser. This information is available in real-time for different frequencies, although this article uses its monthly version. With unrestricted public access, the data are available at the national, departmental, and large city levels.

Google Trends reports an index ranging from 0 to 100 over the selected period. "The numbers reflect search interest relative to the highest value in a region and over time. A value of 100 indicates the maximum popularity of a term, while 50 and 0 indicate popularity that is half or less than 1%, respectively, to the highest value" (Google, 2022).

Table 1 presents the search terms considered. The 23 queries included were associated with the job search process, platforms, tools, or online entities dedicated to job search and/or personnel recruitment, and government aid for unemployment (subsidies). Additionally, the previous study conducted in Colombia by Cardona and Rojas (2017) and the search suggestions highlighted by the Google Trends platform with greater relative importance recorded during the last years were considered.

Table 1. Search Terms in Google Trends for Cali, Medellín, Bogotá D.C., and Popayán

Searched Term		Searched Term	
Empleo	GT_EMP	*Computrabajo*	GT_COMPUT
Ofertas de empleo	GT_OEMP	*Ofertas de empleo*	GT_OTRAB
Vacantes	GT_VAC	*Olx empleo*	GT_OLXE
Hoja de vida	GT_HOJAV	*Trabajo Colombia*	GT_TRABCOL
Trabajo sin experiencia	GT_AGEMP	*Trabajo*	GT_TRAB
Agencia de empleo	GT_TSEXP	*Servicio de empleo*	GT_SEREMP
Sena empleo	GT_SENA	*Familias en acción*	GT_FAMACC
El empleo	GT_EEMP	*Subsidio de desempleo*	GT_SUBDES
Indeed	GT_INDEED	*Prestamos*	GT_PREST

Searched Term		Searched Term	
Bolsa de empleo	GT_BOLEMP	*Retiro cesantías*	GT_RCES
Busco trabajo	GT_BTRA	*Cesantías*	GT_CES
Clasificados	GT_CLAS		

Source:

Variables Selected with ML Techniques

Using the ML techniques of Backward Stepwise and Lasso, 17 variables were selected for Cali and Medellín, 13 from DANE or the Bank of the Republic, and 4 variables related to search terms in Google Trends. For Bogotá, 11 variables were selected, 2 from Google Trends. In Popayán, the two methods selected 31 variables, 22 from DANE or the Bank of the Republic and 9 from Google Trends. Variables were selected with agreement in both methods (see Annex B).

The first interesting result is that the ML techniques confirm the importance of including some of the Google Trends searches in the database for the labor market indicator in combination with traditional employment metrics and variables of gender, age, specific productive sectors, and informality.

Figure 3 shows the graphical correlation between the selected variables, which provides initial information on the comovement that is later estimated with the DFM. In Cali, a strong positive correlation was evidenced between the employment rate and the workforce, employment in the industrial sector, and female participation. Conversely, a strong negative correlation was found between the employment rate and the total unemployment rate, youth unemployment, and female unemployment. Regarding the unemployment rate, there was a positive relationship between the total, youth, and female unemployment rates. It also highlighted that the four search terms from Google Trends positively correlated with employment and unemployment rates.

In Medellín, the behavior was similar, although it was noted that one of the search terms, in this case, "empleo" (employment), showed a negative correlation with the city's total unemployment rate and that the correlation between the four search terms was positive. In Bogotá, the two selected search terms "subsidio desempleo" (unemployment subsidy) and "retiro de cesantías" (withdrawal of severance pay) showed a negative correlation with the employment rate. A positive correlation was evidenced between them.

In Popayán, on the other hand, a significant positive correlation was evidenced between the employment rate, the workforce, female labor participation, and employment, and the personnel employed in the trade, accommodation, and food sectors. Regarding the unemployment rate, the highest positive correlations were recorded with the youth and female unemployment rates. Of the terms searched on Google, the positive correlation between the term "subsidio desempleo" (unemployment subsidies) and the unemployment rate stood out, and the negative correlation between the search term "clasificados" (classifieds) and "subsidio desempleo" (unemployment subsidy).

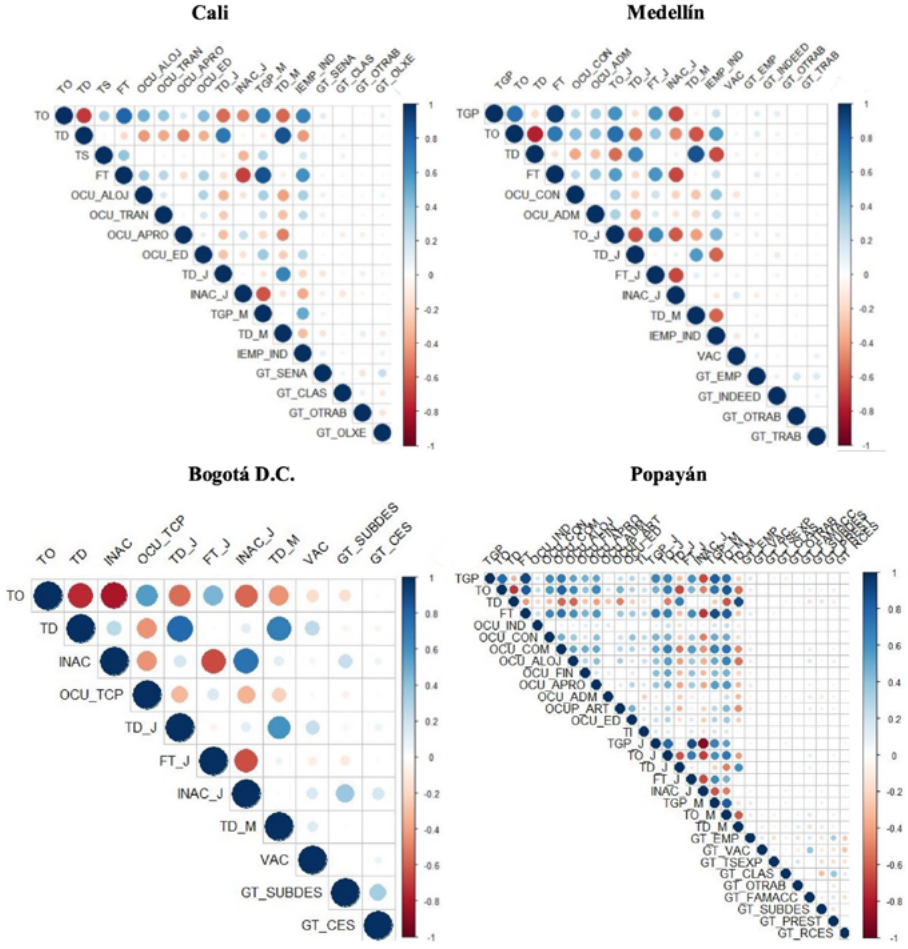

Figure 3. Correlation Between the Selected Variables in Cali, Medellín, Bogotá D.C., and Popayán

Source: Authors' elaboration.

Results

Using the DFM (equation 3), a common factor (F) or comovement among the variables selected with ML techniques for the labor market of the four cities was estimated using the Principal Components (PC) method. Figure 4 shows two principal components and demonstrates that the first principal component explained the largest proportion of the total variance for the four

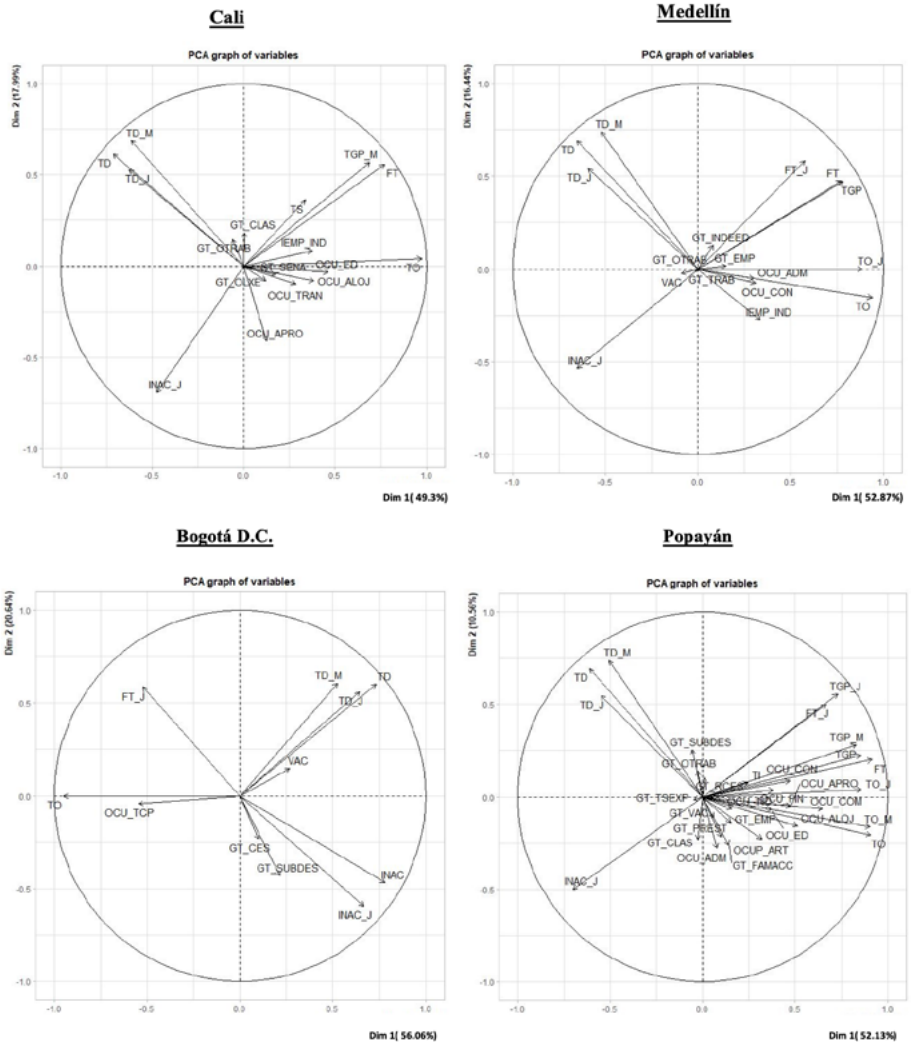

Figure 4. Representation of Variables in the First Two Components (Dim 1 and Dim 2) of the Labor Market Indicator in Cali, Medellín, Bogotá D.C., and Popayán

Source: Authors' elaboration.

regions analyzed. In Cali, the first component accounted for 49.3% of the total variance, while the second accounted for 17.9%. In Medellín, 52.8% and 16.4%; in Bogotá D.C., 56% and 20.6%; and in Popayán, 52.3% and 10.5%, respectively. The labor market indicator was constructed based on the information captured by the first principal component in each city.

Tables 2-5 present the weights (*P*) estimated by the PC method within the DFM (equation 3). These are converted into percentage terms representing each variable's contribution to each city's labor market indicator.

For the city of Cali, 6 out of 17 variables contribute about 76% to the labor market indicator (dim1): employment rate (22.21%), labor force (13.8%), unemployment rate (11.85%), overall participation rate of women (11.12%), youth unemployment rate (9.1%), and women's unemployment rate (8.8%), as part of the labor market indicator in Cali. The four Google Trends variables contribute about 1.1%.

In Medellín, 6 out of 17 variables contribute 74% to the labor market indicator: employment rate (17.6%), youth employment rate (15.6%), overall participation rate (12.2%), labor force (11.9%), unemployment rate (8.5%), and inactive youth (8.4%). Notably, the 4 Google Trends variables contribute 0.6% to the indicator (Table 3).

In Bogotá, 5 out of 11 variables contribute about 75% to the indicator: employment rate (23.1%), population outside the labor force (15.6%), unemployment rate (13.8%), inactive youth (11.4%), and youth unemployment rate (10.7%). The two selected Google Trends variables contribute 1.44% to the indicator.

In Popayán, the largest contributions to the indicator are observed in the employment rate (9.49%), the labor force (9.63%), overall participation rate (8.3%), women's employment rate (9.41%), and youth employment rate (8.45%). The selected Google Trends variables contribute 0.7% to the indicator.

Table 2. Contribution of Variables to the Labor Market Indicator of Cali

No.	Variable	Dim.1
1	Employment rate	22.21
2	Unemployment rate	11.85
3	Underemployment rate	2.66
4	Labor force	13.84
5	Employed accommodation and food services	3.44
6	Employed transportation and warehousing	1.90
7	Employed in professional activities	0.38
8	Employed as domestic employee	4.99
9	Youth unemployment rate	9.11
10	Inactive youth	5.23
11	Overall participation rate of women	11.12
12	Women's unemployment rate	8.80
13	Industrial employment rate	3.30
14	Google search for the word “sena empleo”	0.71
15	Google search for the word “clasificados”	0.00
16	Google search for the word “ofertas de trabajo”	0.08
17	Google search for the word “olx empleo”	0.35

Source: Authors' elaboration.

Table 3. Contribution of Variables to the Labor Market Indicator of Medellín

No.	Variables	Dim1
1	Overall Participation Rate	12.2
2	Employment rate	17.6
3	Unemployment rate	8.5
4	Labor Force	11.9
5	Employed Construction	1.9
6	Employed Public administration, education and health	1.8
7	Youth employment rate	15.6
8	Youth unemployment rate	7.0
9	Youth labor force	6.6
10	Inactive youth	8.4
11	Women's unemployment rate	5.4
12	Industrial employment rate	2.2
13	Job vacancies or job offers according to print ads	0.2
14	Google search for the word “empleo”	0.5
15	Google search for the word “indeed”	0.1
16	Google search for the word “ofertas de trabajo”	0.0002
17	Google search for the word “trabajo”	0.006

Source: Authors' elaboration.

Table 4. Contribution of Variables to the Labor Market Indicator of Bogotá D.C.

No.	Variables	Dim.1
1	Employment rate	23.18
2	Unemployment rate	13.85
3	Population outside the labor force	15.68
4	Employed as self-employed	7.64
5	Youth unemployment rate	10.70
6	Youth labor force	7.05
7	Inactive youth	11.49
8	Women's unemployment rate	7.11
9	Job vacancies or job offers according to print ads	1.86
10	Google search for the word “subsidio desempleo”	1.18
11	Google search for the word “cesantias”	0.26

Source: Authors' elaboration.

Table 5. Contribution of Variables to the Labor Market Indicator of Popayán

No.	Variable	Dim.1
1	Overall participation rate	8.39
2	Employment rate	9.49
3	Unemployment rate	4.28
4	Labor force	9.63
5	Employed industry	0.28
6	Employed construction	2.57
7	Employed trade and repair of vehicles	4.84
8	Employed accommodation and food services	3.03
9	Employed financial and insurance activities	1.67
10	Employed professional activities	2.62
11	Employed public administration, education and health	0.07
12	Occupied Artistic, entertainment and recreation activities	1.17
13	Employed as domestic employee	1.58
14	Informality rate	0.69
15	Overall youth participation rate	6.08
16	Youth employment rate	8.45
17	Youth unemployment rate	3.47
18	Youth Labor Force	5.10
19	Inactive youth	5.64
20	Overall Participation Rate Women	7.88
21	Women's employment rate	9.41
22	Women's unemployment rate	2.94
23	Google search for the word “empleo”	0.27
24	Google search for the word "vacantes"	0.04
25	Google search for the word “agencia de empleo”	0.03
26	Google search for the word “clasificados”	0.01
27	Google search for the word “ofertas de trabajo”	0.01
28	Google search for the word “familias en accion”	0.23
29	Google search for the word “subsidio desempleo”	0.04
30	Google search for the word “prestamos”	0.11
31	Google search for the word “ retiro cesantias"	0.00

Source: Authors' elaboration.

Figure 5.A compares the trajectory of the common factor estimated for the labor market of each city and the employment rate (standardized monthly variations). It is evident that the time series of the employment rate adjusted to the first principal component. A match in the sign of most of the monthly variations was observed, as well as similarity in the periods of greatest expansion and contraction, including the time of the pandemic. However, as expected, the trajectories are not identical because the indicator contains additional information on the labor market that goes beyond employment.

Similarly, Figure 5.B presents the relationship between the estimated labor market indicator and the regional unemployment rate. The patterns observed reinforce the indicator's robustness, showing an inverse relationship with the unemployment rate, as expected. This inverse correlation is particularly pronounced during the pandemic, when a sharp decline in the labor market indicator coincides with a significant peak in unemployment in the cities analyzed.

To provide a clearer and more interpretable trend, Figure 6 presents the indicator (estimated common factor) smoothed with the Hodrick-Prescott filter, rescaled and transformed into an index with February 2020 = 100 (pre-pandemic level). An increase (decrease) in the indicator suggests an improvement (worsening) of the regional labor market, integrating traditional employment metrics and social and equity variables, expectations, and specific productive sectors into this signal.

The indicator also shows greater heterogeneity among regional labor markets until 2014 and a convergence process from 2015. The impact of the pandemic on the labor market is visible in all four cities. Popayán was the city that registered the greatest deterioration compared to the pre-pandemic level: the indicator was 32% below the levels recorded in February 2020. Cali is the second most affected city, with an indicator 29.2% below the pre-pandemic level. In Bogotá and Medellín, the indicator remained 25.2% and 17.6% below the February 2020 levels, respectively.

Since the last months of 2020, the labor market indicator reports a recovery trend that was briefly interrupted by the national strike and social protests during May and June 2021 in the cases of Cali and Popayán. In 2022, the indicator reports the continuation of the recovery. Medellín is the first city to surpass pre-pandemic levels in February 2022. Cali's labor market, for its part, surpassed pre-pandemic levels in June 2022, and Popayán achieved it in December 2022. In contrast, the labor market in Bogotá D.C. still in December 2022 remained 6.2% below pre-pandemic conditions.

Table 6 summarizes the differences in the signals that could be obtained when comparing regional labor markets using only the information from the unemployment rate and the employment rate (see section 3) and using the labor market indicator. The similarities and differences appear when comparing traditional employment metrics with an indicator that allows for a more comprehensive understanding of the labor market situation.

For the practical application and monthly update of the estimated labor market indicator, a publication delay of approximately 45 to 50 days is expected, considering the availability of its component variables (see Annex A).

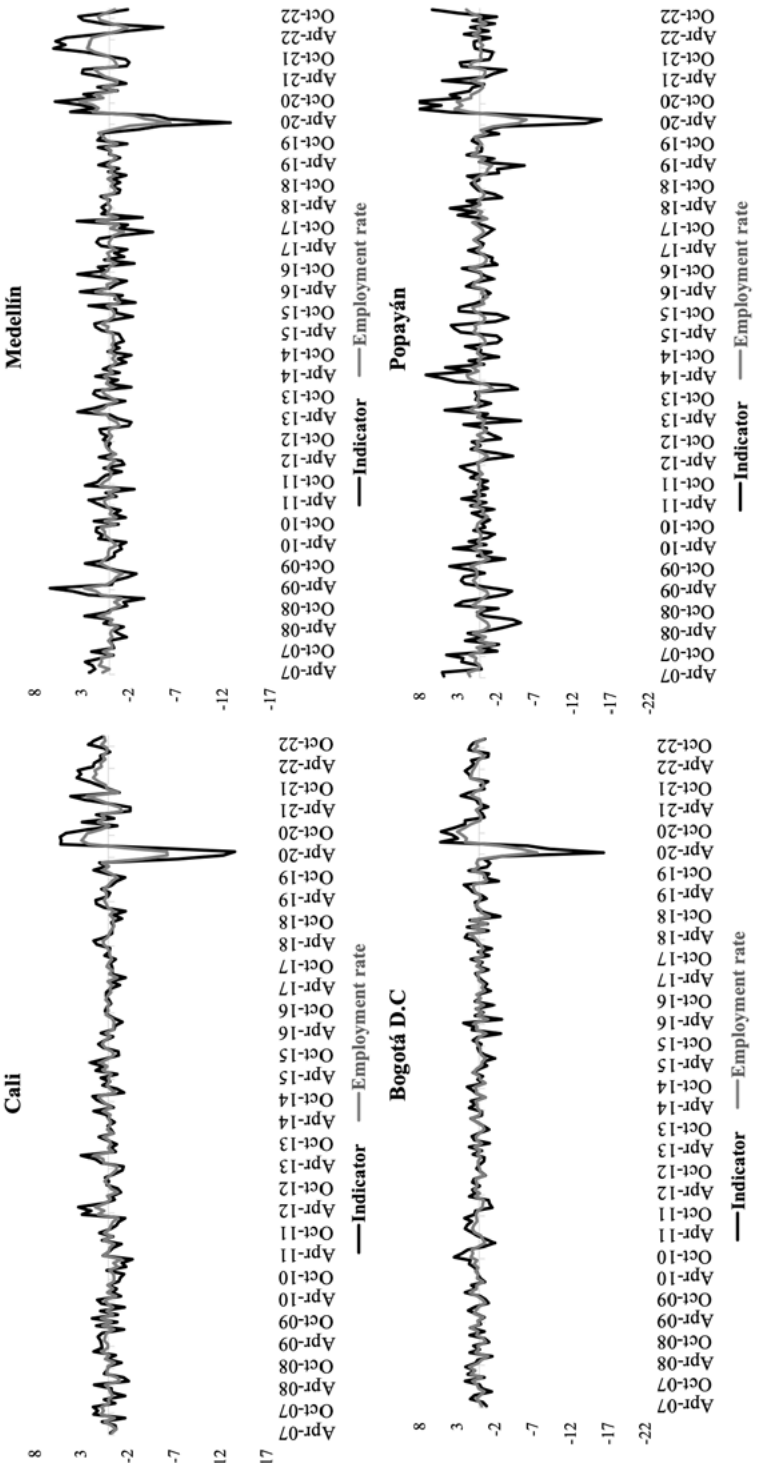

Figure 5.A. Labor Market Indicator and Regional Employment Rate–(Standardized Monthly Variation)

Source: Authors' elaboration.

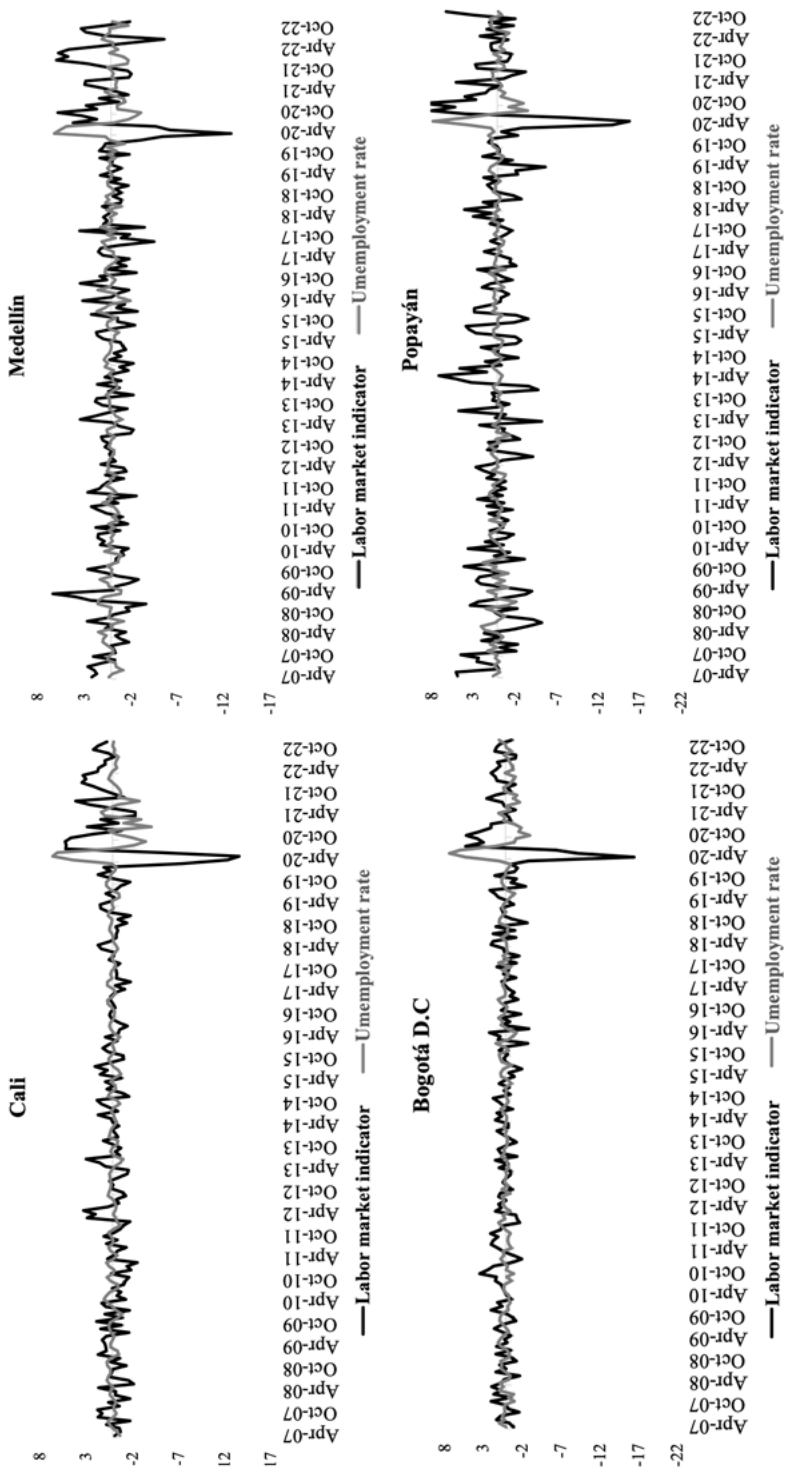

Figure 5.B. Labor Market Indicator and Regional Umemployment Rate–(Standardized Monthly Variation)

Source: Authors' elaboration.

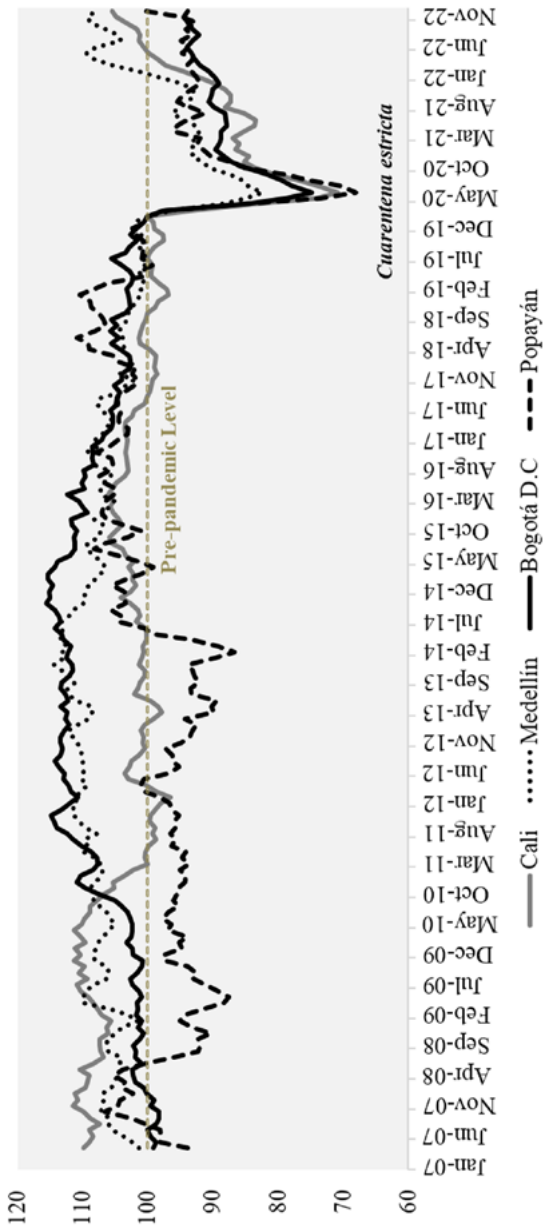

Figure 6. Trajectory of the Regional Labor Market Indicator (2007-2022)
(Index February 2020=100, Pre-pandemic Level)
Source: Authors' elaboration.

Table 6. Summary of Compared Signals on the Labor Market

Métrica	Labor Market Conditions Before the Pandemic Impact (Feb20)	Impact of the Pandemic in 2020	Impact of the National Strike in 2021	Recovery in 2022
Regional Labor Market Indicator	Medellín recorded the best performance, and Cali the worst.	Popayán and Cali suffered the greatest impact, while Medellín experienced the least.	Popayán and Cali suffered the greatest impact, while Medellín experienced the least.	Popayán and Bogotá D.C. are the most lagging cities, while Medellín and Cali have the greatest recovery.
Unemployment Rate	Bogotá D.C. recorded the best performance, and Cali the worst.	Popayán and Medellín suffered the greatest impact, while Bogotá experienced the least.	Popayán and Cali suffered the greatest impact, while Medellín experienced the least.	All cities are converging towards recovery.
Employment Rate	Bogotá D.C. recorded the best performance, and Medellín the worst.	Popayán and Cali suffered the greatest impact, while Bogotá experienced the least.	Popayán and Cali suffered the greatest impact, while Bogotá experienced the least.	Popayán is the most lagging city, while Bogotá, Cali, and Medellín stand out for their resilience.

Source: Authors' elaboration.

Conclusions

This article proposes a methodology to estimate an indicator that summarizes the aggregated evolution of the labor market by combining economic, social, inequality, and expectation variables. This allows for a standardized and more consistent comparison among regional labor markets. The methodology is applied in four Colombian cities (Cali, Medellín, Bogotá D.C., and Popayán). It covers the period of the Covid-19 pandemic impact and the subsequent evolution, a time when several authors have emphasized the increase in labor market heterogeneities.

The methodology employs Machine Learning techniques to select the most relevant variables for estimating the indicator. Once the comovement among them is estimated, the selected variables allow for a synthetic characterization of labor markets. For this purpose, a Dynamic Factor Model is used, which allows aggregating the variables based on appropriately estimated weights (contributions).

The Machine Learning techniques confirmed the importance of including Google Trends searches in the labor market indicator in combination with traditional employment metrics and variables of gender, age, specific productive sectors, and informality. Google searches related to the labor market allow considering public perceptions and expectations regarding employment and integration of the indicator within the emerging trend of data analysis using information from social networks and the Internet.

The impact of the pandemic on the labor market is visible in all four cities, as well as the heterogeneities during the recovery. The indicator's trajectory coincides with the signals reported by the employment rate and the unemployment rate but also presents differences because it contains additional information on gender, age, informality, productive sectors, and Google Trends data, allowing for a more comprehensive understanding of the labor market situation.

References

Adams-Prassl, A., Boneva, T., Golin, M., & Rauh, C. (2021). Inequality in the impact of the coronavirus shock: Evidence from real-time surveys. *Journal of Public Economics*, 189, 104-245.

Alon, T., Doepke, M., Olmstead-Rumsey, J., & Tertilt, M. (2021). *The Impact of COVID-19 on Gender Equality* [NBER Working Paper Series, No. 26947].

Bonilla, L., & Gaviria, C. (2020). El mercado laboral colombiano en tiempos de la pandemia. *Análisis de Coyuntura, 2*(4), 1-19.

Baldacci, E., Marcellino, M., Papailias, F., Kapetanios, G., Buono, D., Krische, S., & Mazzi, G. (2016). *Big Data and Macroeconomic Nowcasting: From data access to modelling*. Eurostat. https://doi.org/10.2785/3605875

Cajner, T., Crane, L., Decker, R., Grigsby, J., Hamins-Puertolas, A., Hurst, E., Kurz, C., & Yildirmaz, A. (2021). The U.S. labor market during the beginning of the pandemic recession. *Journal of Public Economics*, 193, 104-312.

Chung, H., Fallick, B., Nekarda, C., y Ratner, D. (2015). *Assessing the Change in Labor Market Conditions* [Working Papers–Old Series, N.º 1438). Federal Reserve Bank of Cleveland. https://ideas.repec.org/p/fip/fedcwp/1438.html

Choi, J., & Shin, D. (2019). The roles of differencing and dimension reduction in machine learning forecasting of employment level using the FRED big data. *Communications for Statistical Applications and Methods, 26*(5), 497-506. https://doi.org/10.29220/CSAM.2019.26.5.497

Chen, J., & Wu, J. (2023). The prediction of Chongqing's GDP based on the LASSO method and chaotic whale group algorithm–back propagation neural network–ARIMA model. *Scientific Reports*, *13*(1), 15002.

Choi, H. & H. Varian. 2012. Predicting the Present with Google Trends, *Economic Record*, *88*(s1): 29.

Caperna, G., Colagrossi, M., Geraci, A., & Mazzarella, G. (2020). Googling unemployment during the pandemic: Inference and nowcast using search data (N.º 2020/04). JRC Working Papers in Economics and Finance.

Cardona, L., & Rojas, J. (2017). *Pronósticos para la tasa de desempleo en Colombia a partir de Google Trends* [N.º 016050]. Departamento Nacional de Planeación.

Campos Vázquez, R., & López-Araiza, S. (2020). Grandes datos, Google y desempleo. *Estudios Económicos*, *35*(1), 125-151.

Chung, H., Fallick, B., Nekarda, C., & Ratner, D. (2015). *Assessing the Change in Labor Market Conditions* [Working Papers–Old Series, N.º 1438]. Federal Reserve Bank of Cleveland. https://ideas.repec.org/p/fip/fedcwp/1438.html

Dvorkin, M. (2015). Assessing the Health of the Labor Market: The Unemployment Rate vs. Other Indicators. *The Regional Economist*. https://ideas.repec.org/a/fip/fedlre/00064.html

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

Google (2022). *Google Trends: entendiendo los datos*. https://newsinitiative.withgoogle.com/resources/trainings/google-trends-understanding-the-data/

Hakkio, C., & Willis, J. (2013). Assessing labor market conditions: The level of activity and the speed of improvement. *Macro Bulletin*, July 18. https://ideas.repec.org/a/fip/fedkmb/y2013ijuly18.html

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). Springer.

Jokubaitis, S., Celov, D., & Leipus, R. (2021). Sparse structures with LASSO through principal components: Forecasting GDP components in the short run. *International Journal of Forecasting*, *37*(2), 759-776.

Jun, S., Yoo, H., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, 130, 69-87.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jolliffe, I. (2002). *Principal component analysis for special types of data* (pp. 338-372). Springer.

Jolliffe, I., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, (374). https://doi.org/10.1098/rsta.2015.0202

Morales, L., Mejía, L., Pulido, J., Flórez, L., Valderrama, F., Hermida, D., & Mahecha, K. (2022). Efectos de la pandemia por Covid-19 en el mercado laboral colombiano. In *Covid-19: Consecuencias y desafíos en la economía colombiana* (pp. 63-86). Banco de la República de Colombia.

Medeiros, M., Vasconcelos, G., Veiga, Á., & Zilberman, E. (2019). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, 1-22. https://doi.org/10.1080/07350015.2019.1637745

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Miller, A. J. (2002). "Subset Selection in Regression." Chapman and Hall/CRC.

Orozco-Gallo, A., Vidal-Alejandro, P., Sanabria-Domínguez, J., & Collazos-Rodríguez, J. (2021). Indicador coincidente de actividad económica en la recesión pandémica: el caso del Caribe colombiano. *Documento sobre economía regional y urbana; No. 298*.

Orozco-Castañeda, J. M., Sierra-Suárez, L. P., & Vidal, P. (2024). Labor market forecasting in unprecedented times: A machine learning approach. *Bulletin of Economic Research*, *76*(4), 893-915.

Qin, D., Van Huellen, S., Wang, Q., & Moraitis, T. (2022). Algorithmic modelling of financial conditions for macro predictive purposes: pilot application to USA data. *Econometrics, 10*(2), 22.

Ramos-Veloza, M., Cristiano-Botia, D., & Hernandez-Bejarano, M. (2021). Labor Market Indicator for Colombia. *Latin American economic review*, *30*, 1–32. https://doi.org/10.47872/laer.v30.17

Sierra-Suárez, L., Collazos-Rodríguez, J., Sanabria-Domínguez, J., & Vidal-Alejandro, P. (2017). La construcción de indicadores de la actividad económica: una revisión bibliográfica. *Apuntes del CENES*, *36*(64), 79-107.

Sierra-Suárez, L. P., Alejandro, P. V., & Cerón, J. (2022). Una mirada regional al impacto económico del Covid-19 desde el indicador mensual de actividad económica (IMAE) para el Valle del Cauca. In *Covid-19 consecuencias y desafíos en la economía colombiana: Una mirada desde las universidades* (pp. 289-303). Universidad del Rosario.

Swanson, N., & Xiong, W. (2018). Big data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal*

of Economics/Revue canadienne d'économique, *51*(3), 695-746. https://doi.org/10.1111/caje.12336

Simionescu, M., & Raišienė, A. (2021). A bridge between sentiment indicators: What does Google Trends tell us about COVID-19 pandemic and employment expectations in the EU new member states? *Technological Forecasting and Social Change*, *173*, 121170.

Stock, J., & Watson, M. (2012). Disentangling the Channels of the 2007-2009 Recession (N.º w18094). National Bureau of Economic Research.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Velasco, J. (2021). Los impactos de la pandemia de la Covid-19 en los mercados laborales de América Latina. Compendium: Cuadernos de Economía y Administración, *8*(2), 99-120.

Vidal, P., Sierra, L., Sanabria, J., & Collazos, J. (2017). A Monthly Regional Indicator of Economic Activity: An Application for Latin America. *Latin American Research Review*, *52*(4), 589-605.

Varian, H. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, *28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Zmitrowicz, K., & Khan, M. (2014). Beyond the Unemployment Rate: Assessing Canadian and U.S. Labour Markets Since the Great Recession. *Bank of Canada Review*, 2014 (Spring), 42-53.

Annexes

Annex A. Database for Cali, Medellín, Bogotá D.C., and Popayán

No.	Variable	acronym	Unit of measure	Publication lag	Source	City available
1	Overall Participation Rate	TGP	Percentage-Rate	31 days	DANE-GEIH	4 cities
2	Employment rate	TO	Percent-Rate	31 days	DANE-GEIH	4 cities
3	Unemployment rate	TD	Percent-Rate	31 days	DANE-GEIH	4 cities
4	Underemployment rate	TS	Percent-Rate	31 days	DANE-GEIH	4 cities
5	Labor force	FT	Thousands of people	31 days	DANE-GEIH	4 cities
6	Population outside the labor force	INAC	Thousands of people	31 days	DANE-GEIH	4 cities
7	Employed Agriculture	OCU_AGR	Thousands of people	31 days	DANE-GEIH	4 cities
8	Employed Industry	OCU_IND	Thousands of people	31 days	DANE-GEIH	4 cities
9	Employed Construction	OCU_CON	Thousands of people	31 days	DANE-GEIH	4 cities
10	Employed Trade and repair of vehicles	OCU_COM	Thousands of people	31 days	DANE-GEIH	4 cities
11	Employed Accommodation and food services	OCU_ALOJ	Thousands of people	31 days	DANE-GEIH	4 cities
12	Employed Transportation and storage	OCU_TRAN	Thousands of people	31 days	DANE-GEIH	4 cities
13	Occupied Financial and insurance activities	OCU_FIN	Thousands of people	31 days	DANE-GEIH	4 cities
14	Employed Professional activities	OCU_APRO	Thousands of people	31 days	DANE-GEIH	4 cities
15	Occupied Public administration, education and health	OCU_ADM	Thousands of people	31 days	DANE-GEIH	4 cities
16	Occupied Artistic, entertainment and recreation activities	OCUP_ART	Thousands of people	31 days	DANE-GEIH	4 cities
17	Employed as domestic employee	OCU_ED	Thousands of people	31 days	DANE-GEIH	4 cities
18	Employed as self-employed	OCU_TCP	Thousands of people	31 days	DANE-GEIH	4 cities
19	Informality rate	TI	Percentage-Rate	45 days	DANE-GEIH	4 cities
20	Overall Youth Participation Rate	TGP_J	Percentage-Rate	45 days	DANE-GEIH	4 cities
21	Youth Employment Rate	TO_J	Percentage-Rate	45 days	DANE-GEIH	4 cities
22	Youth unemployment rate	TD_J	Percentage-Rate	45 days	DANE-GEIH	4 cities
23	Youth Labor Force	FT_J	Thousands of people	45 days	DANE-GEIH	4 cities
24	Inactive youth	INAC_J	Thousands of people	45 days	DANE-GEIH	4 cities
25	Overall Participation Rate Women	TGP_M	Percent-Rate	45 days	DANE-GEIH	4 cities
26	Women's employment rate	TO_M	Percent-Rate	45 days	DANE-GEIH	4 cities
27	Women's unemployment rate	TD_M	Percentage-Rate	45 days	DANE-GEIH	4 cities
28	Labor force Women	FT_M	Thousands of people	45 days	DANE-GEIH	4 cities
29	Industrial employment rate	IEMP_IND	Index	45 days	DANE-GEIH	Cali, Bogotá y Medellín
30	Job vacancies or job offers according to print advertisements	VAC	No. of vacancies	30 days	Bank of the Republic of Colombia	4 cities
31	Google search for the word "empleo"	GT_EMP	Search rate	1 day	Google Trends	4 cities
32	Google search for the word "ofertas de empleo"	GT_OEMP	Search Index	1 day	Google Trends	4 cities
33	Google search for the word "vacantes"	GT_VAC	Search index	1 day	Google Trends	4 cities
34	Google search for the word "hoja de vida"	GT_HOJAV	Search index	1 day	Google Trends	4 cities
35	Google search for the word "agencia de empleo"	GT_AGEMP	Search index	1 day	Google Trends	4 cities
36	Google search for the word "trabajo sin experiencia"	GT_TSEXP	Search index	1 day	Google Trends	4 cities
37	Google search for the word "sena empleo"	GT_SENA	Search index	1 day	Google Trends	4 cities
38	Google search for the word "el empleo"	GT_EEMP	Search index	1 day	Google Trends	4 cities
39	Google search for the word "indeed"	GT_INDEED	Search index	1 day	Google Trends	4 cities
40	Google search for the word "bolsa de empleo"	GT_BOLEMP	Search index	1 day	Google Trends	4 cities
41	Google search for the word "busco trabajo"	GT_BTRA	Search index	1 day	Google Trends	4 cities
42	Google search for the word "clasificados"	GT_CLAS	Search index	1 day	Google Trends	4 cities
43	Google search for the word "computrabajo"	GT_COMPUT	Search index	1 day	Google Trends	4 cities
44	Google search for the word "ofertas de trabajo"	GT_OTRAB	Search index	1 day	Google Trends	4 cities
45	Google search for the word "olx empleo".	GT_OLXE	Search index	1 day	Google Trends	4 cities
46	Google search for the word "trabajo colombia"	GT_TRABCOL	Search index	1 day	Google Trends	4 cities
47	Google search for the word "trabajo	GT_TRAB	Search index	1 day	Google Trends	4 cities
48	Google Search for the word "servicio de empleo"	GT_SEREMP	Search index	1 day	Google Trends	4 cities
49	Google search for the word "familias en accion"	GT_FAMACC	Search index	1 day	Google Trends	4 cities
50	Google search for the word "subsidio desempleo"	GT_SUBDES	Search index	1 day	Google Trends	4 cities
51	Google search for the word "prestamos"	GT_PREST	Search index	1 day	Google Trends	4 cities
52	Google search for the word " retiro cesantias"	GT_RCES	Search index	1 day	Google Trends	4 cities
53	Google search for the word "cesantias"	GT_CES	Search index	1 day	Google Trends	4 cities

Annex B. Variables Selected by the Lasso and Backward Methods

No.	Variables	acronym	Cali	Medellín	Bogotá	Popayán
1	Overall Participation Rate	TGP	û	ü	û	ü
2	Employment rate	TO	ü	ü	ü	ü
3	Unemployment rate	TD	ü	ü	ü	ü
4	Underemployment rate	TS	ü	û	û	û
5	Labor force	FT	ü	ü	û	ü
6	Population outside the labor force	INAC	û	û	ü	û
7	Employed Agriculture	OCU_AGR	û	û	û	û
8	Employed Industry	OCU_IND	û	û	û	ü
9	Employed Construction	OCU_CON	û	ü	û	ü
10	Employed Trade and repair of vehicles	OCU_COM	û	û	û	ü
11	Employed Accommodation and food services	OCU_ALOJ	ü	û	û	ü
12	Employed Transportation and storage	OCU_TRAN	ü	û	û	û
13	Occupied Financial and insurance activities	OCU_FIN	û	û	û	ü
14	Employed Professional activities	OCU_APRO	ü	û	û	ü
15	Occupied Public administration, education and health	OCU_ADM	û	ü	û	ü
16	Occupied Artistic, entertainment and recreation activities	OCUP_ART	û	û	û	ü
17	Employed as domestic employee	OCU_ED	ü	û	û	ü
18	Employed as self-employed	OCU_TCP	û	û	ü	û
19	Informality rate	TI	û	û	û	ü
20	Overall Youth Participation Rate	TGP_J	û	û	û	ü
21	Youth Employment Rate	TO_J	û	ü	û	ü
22	Youth unemployment rate	TD_J	ü	ü	ü	ü
23	Youth Labor Force	FT_J	û	ü	ü	ü
24	Inactive youth	INAC_J	ü	ü	ü	ü
25	Overall Participation Rate Women	TGP_M	ü	û	û	ü
26	Women's employment rate	TO_M	û	û	û	ü
27	Women's unemployment rate	TD_M	ü	ü	ü	ü
28	Labor force Women	FT_M	û	û	û	û
29	Industrial employment rate	IEMP_IND	ü	ü	û	
30	Job vacancies or job offers according to print advertisements	VAC	û	ü	ü	
31	Google search for the word "empleo"	GT_EMP	û	ü	û	ü
32	Google search for the word "ofertas de empleo"	GT_OEMP	û	û	û	û
33	Google search for the word "vacantes"	GT_VAC	û	û	û	ü
34	Google search for the word "hoja de vida"	GT_HOJAV	û	û	û	û
35	Google search for the word "agencia de empleo"	GT_AGEMP	û	û	û	û
36	Google search for the word "trabajo sin experiencia"	GT_TSEXP	û	û	û	ü
37	Google search for the word "sena empleo"	GT_SENA	ü	û	û	û
38	Google search for the word "el empleo"	GT_EEMP	û	û	û	û
39	Google search for the word "indeed"	GT_INDEED	û	ü	û	û
40	Google search for the word "bolsa de empleo"	GT_BOLEMP	û	û	û	û
41	Google search for the word "busco trabajo"	GT_BTRA	û	û	û	û
42	Google search for the word "clasificados"	GT_CLAS	ü	û	û	ü
43	Google search for the word "computrabajo"	GT_COMPUT	û	û	û	û
44	Google search for the word "ofertas de trabajo"	GT_OTRAB	ü	ü	û	ü
45	Google search for the word "olx empleo".	GT_OLXE	ü	û	û	û
46	Google search for the word "trabajo colombia"	GT_TRABCOL	û	û	û	û
47	Google search for the word "trabajo	GT_TRAB	û	ü	û	û
48	Search in Google for the word "servicio de empleo"	GT_SEREMP	û	û	û	û
49	Google search for the word "familias en accion"	GT_FAMACC	û	û	û	ü
50	Google search for the word "subsidio desempleo"	GT_SUBDES	û	û	ü	ü
51	Google search for the word "prestamos"	GT_PREST	û	û	û	ü
52	Google search for the word " retiro cesantias"	GT_RCES	û	û	û	ü
53	Google search for the word "cesantias"	GT_CES	û	û	ü	û
Total number of Google Trends variables selected for inclusion in the indicator			4	4	2	9
Total number of variables selected for inclusion in the indicator			17	17	11	31