Borradores de ECONOMÍA

An Exploratory Analysis of Heterogeneity on Regional Labour Markets and Unemployment Rates in Colombia: An MFACT approach

Por: Camilo Alberto Cárdenas Hurtado, María Alejandra Hernández Montes, Jhon Edwar Torres Gorron

Núm. 802
2014

BANCO DE LA REPÚBLICA
COLOMBIA

An Exploratory Analysis of Heterogeneity on Regional Labour Markets and Unemployment Rates in Colombia: An MFACT approach*

Camilo Alberto Cárdenas Hurtado†
María Alejandra Hernández Montes‡
Jhon Edwar Torres Gorron§

Abstract

In this paper we study the structural determinants of differentials in unemployment rates and labour markets' performance for colombian cities. Following the framework proposed by Elhorst (2003) and using cross-sectional data for 23 metropolitan areas, we apply an extension of a principal axes method proposed by Bécue-Bertaut and Pagès (2004, 2008), Multiple Factor Analysis for Multiple Contingency Tables (MFACT), in order to establish unobserved factors that are relevant when disentangling the heterogeneity captured by groups of variables that are considered to explain regional unemployment differentials. Our findings suggest that differences on qualified labour supply levels, participation incentives and age structure are important to understand regional heterogeneity on labour markets and unemployment rates. In addition, we find that cities that display high unemployment rates do not necessarily share the same characteristics, that is, frictions that originate unemployment are not the same across colombian cities.

Keywords: Unemployment rate, regional, heterogeneity, differentials, factor analysis.

JEL: R23, J40.

1 Introduction

Both high levels and persistence of unemployment rates, as well as the complex dynamics observed on labour market structures in Colombia, have puzzled local economists for decades now. Although some queries have been studied over the past few years (see Urrutia, 2001; Arango and Hamann, 2013; among others as an starting reference), there are still several unanswered questions that, if solved, might lead to a better understanding of the convoluted particularities of labour

*The opinions, statements, findings and interpretations presented in this paper are responsibility of the authors and do not represent those of Banco de la República nor of its Board of Directors. Usual additional disclaimers apply. We thank Daniel Quintero Castro (dquintca@banrep.gov.co), who participated actively on the early stages of this paper. Comments from Luis Eduardo Arango, Adolfo Cobo, Sebastián Amador and Carmiña Vargas were very helpful, appreciated and acknowledged. Valuable research assistance was received from Jackeline Piraján and Natalia Solano.

†Corresponding author. e-mail: ccardehu@banrep.gov.co
‡e-mail: mhernamo@banrep.gov.co.
§e-mail: jtorrego@banrep.gov.co.

market institutions in our country.

One of the most unexplored topics in colombian labour market literature is regional unemployment, as stated by Arango (2013). Some pioneer works on explaining regional and urban unemployment in Colombia are those of Jaramillo et al. (2000), Galvis (2002), Gamarra (2005) or Barón (2013). However, this topic has not been fully explored by colombian economists and there is still a lot of topics to unveil. Arango (2013) points out that there are noticeable differences between colombian cities when analysing labour markets performance over the past few decades. His findings show that there is an evident heterogeneity between cities on labour market indicators such as unemployment rate, participation rate, occupation rate, underemployment rates, salaries and education.

He shows that some cities, such as Pereira, Popayán and Quibdó, have persistently displayed high unemployment rates over the past few years; while others, like Bogotá, Barranquilla, Bucaramanga and Cali, have seemingly performed better over the same time span. There are several *feasible* explanations for these differences, but still not a single definite one. This article aims to explore such differences by analysing the determinants of regional differentials on unemployment rates, following the framework proposed by Elhorst (2003). We build a high dimensional data set for colombian cities and find the structural determinants that lead to the regional heterogeneity on labour market indicators described by Arango. To our best knowledge, no article has documented what are the driving factors that determine the contrasts in unemployment rates between regions or cities in Colombia .

Since our main goal is descriptive in essence, exploratory multivariate statistical analysis tools are suitable methods to synthesize all the information encoded in a high dimensional dataset into a lower dimensional space that admits a graphical representation. Therefore, we rely on the Multiple Factor Analysis (MFA, Escofier and Pagès 2008) and its extension to a table containing various frequency tables (MFACT), introduced by Bécue-Bertaut and Pagès (2004, 2008). The main characteristics of the methodology are explained in detail. We are also interested in knowing if cities can be grouped into different clusters that share common structural determinants of regional unemployment differentials.

This article consists of six sections, being this introduction the first one. In Section 2 we describe the theoretical and empirical determinants of differentials on regional labour markets proposed by Elhorst (2003), enriched by a complementary literature review. Third section describes the statistical methodology used in this paper and the data. Section 4 covers the main results of the MFACT exercise. Clustering results are shown in Section 5. Last section concludes and suggest that in order to reduce unemployment rates and assure better labour conditions for working age population on colombian cities, it is important to count for the heterogeneity observed on regional labour markets. Our results also suggest that unemployment is the result of several different frictions on labour markets that should be faced from different approaches.

2 Explaining regional labour market differentials

Following Elhorst (2003), variables that explain differentials in regional unemployment fall into one or more of the categories here presented. On one hand, there are endogenous variables that are related to the city's population and the dynamics of regional labour markets; on the other, there are exogenous variables that are not directly related to the labour force nor the equilibrium reaching mechanism. We stress that no attempt is made to be exhaustive in reviewing the existing

literature, since it is not the main goal of this paper. Instead, we focus on representative and influential papers on regional unemployment topics that have enriched labour economics literature over the past few decades. Accordingly, Elhorst states that variables can be categorized into one of the following groups:

Demographic structure

Variables like birth rate, age structure and other related demographic indicators have been found to be determinant on the labour supply size in the long run (Biffl, 1998; Lerman and Schmidt, 1999; Chawla et al., 2007). A region will display persistence on its unemployment rate if its population growth is higher than the employment creation rate. In addition, when the age structure of the population is skewed towards young and old individuals, the region is more likely to display high unemployment rates. (Lottman, 2012).

Participation

Mixed results have been found when assessing the significance of these kind of variables at explaining regional unemployment differentials. It is common to think of a positive (probably non-linear) relationship between unemployment and participation rates. However, it has also been found that higher unemployment rates are usually companied by low participation rates. Several explanations arise: according to Fleisher and Rhodes (1976), low participation rates might reflect low levels of human capital investment and low levels of working life commitment. Also, lower female participation rates are often explained by the presence of children in the household. The latter implies a trade off for female workforce between having a family and pursuing a career (Martínez, 2013). Finally, changes in participation rates greater than those in occupation rates might also yield higher unemployment levels (Blundell and MaCurty, 1999; Da Rocha and Fuster, 2006).

Migration

Immigrants flows reinforce the effects reported for participation variables. If high net migration rates are registered for a certain city or region, the effects on the participation rates might lead to a higher unemployment rate as well. Immigrant flows have been found to be correlated with regional disparities on economic performance and labour market conditions, as stated by Pissarides and Wadsworth (1989) and Blanchard and Katz (1992). However, the effect depends heavily upon the endowments (both human and physical capitals) of the incoming population: If high, demand for qualified workforce is likely to increase, as net investment rates and aggregate productivity will also tend to rise, as in Eggert et al. (2010) and Moretti (2012). If low, however, new inhabitants will enter low skilled unemployment lines, as demand for this type of labour might not increase as fast as supply does (Walden, 2012). For the colombian case, Barón (2013) has reported workforce mobility across departments over the past few years, motivated mainly by economic differences between regions.

Commuting

Commuting costs are the result of the recent *suburbanisation* process observed in most cities or regions around the world and also due to the lack of efficient transport systems (relevant on developing countries). Détang-Dessendre and Gaigné (2009) found that long travelling times and large distances between firms and suburban centres have significant effects on unemployment duration and labour market mismatching. In addition, firms' hiring marketpower is higher when workers incur on high commuting costs, measured by both time and money spending, as argued by Brueckner et al. (2002).

WAGES

Theoretically, higher wages usually have a positive effect on labour supply and a negative effect on labour demand, and in frictionless models, wages are the result of the labour market equilibrium reaching mechanism (see Cahuc and Zylberberg (2004, Ch. 5-7) for a comprehensive approach). Wage differentials across regions have also been understood as a consequence of mobility frictions for workforce between regions or cities (Bande et al., 2008). Also, wages serve as a productivity measure: differentials in wages across regions can occur due to differences in labour productive skills (Burdett and Mortesen, 1998).

REGIONAL GROWTH

Regions with good economic performance usually display low unemployment rates (even structural unemployment) and high productivity indicators. This result can be encompassed in the well known Okun's law framework (Okun, 1962), but at a regional level, as in Oberst and Oelgemollër (2013).

MARKET POTENTIAL

Location factors matter for labour market dynamics: firms tend to settle on regions where there are sales growth potential and stable household consumption perspectives, among other reasons (Krugman, 1995). As a consequence, unemployment rates will be lower in those regions. In addition, some approaches have argued that innovation plays a key role in unemployment reduction. Innovate sectors attract skilled labour force and have multiplier effects on employment in other sectors (Moretti, 2010, 2012).

ECONOMIC STRUCTURE

Regions with a diversified productive structure may be less affected by sector-specific shocks and therefore, will exhibit lower unemployment rates along the business cycle, as argued by Malizia and Ke (1993), Izraeli and Murphy (2003) and Tran (2011). This fact has been widely tested in empirical research, as shown in Gupta (1975), Lottman (2012) and Walden (2012).

ECONOMIC AND SOCIAL BARRIERS

These are unobservable economic and social variables that discourage workforce mobility between regions or cities and therefore, act as frictions in regional labour markets (Elhorst, 2003). Frictions on real-state markets, welfare and social security programs, and general tightness of labour markets are some variables in this group. Walden (2012) and Lottman (2012) provide some recent empirical evidence on this topic.

EDUCATION

Higher educational attainment levels lower the risk of unemployment, rise the likelihood of higher wages and promote labour mobility between regions (Mincer, 1991). Also it has been empirically tested that high levels of human capital stocks have spillover effects over non-educated population on labour market outcomes (Winters, 2013). Although overall quality of workforce skills can not be entirely measured by the average number of scholar years, it is a sufficient indicator that has been found to be negatively correlated with the unemployment rate, even at the regional level (Eggert et al., 2010).

UNIONISATION

From a theoretical point of view, unions' bargain power has been treated as a distortion that deviates labour market from its competitive equilibrium (Cahuc and Zylberberg, 2004, Ch.7). Unionisation has been found to be correlated with lower labour demand dynamics in unionised

sectors and also to be an influential variable on the wage setting mechanism, as argued by Mincer (1981), Lewis (1986), and Farber (1986). More recently, unionisation role in labour market has been explored by Albagli et al. (2004), Freeman (2009) and Krusell and Rudanko (2013).

REGIONAL NATURAL UNEMPLOYMENT RATE AND PERSISTENCE

Some authors argue that regional unemployment rates differences arise due to the persistence and lack of convergence between regional labour markets. The natural rate hypothesis states that unemployment can only be deviated from its long run level if structural changes happen, such as in the market share composition or productivity shocks. This approach have been often treated as a purely statistical problem, thus it has been widely explored on empirical studies such as in Brunello et al. (2000), Gomes and da Silva (2009), Lanzafame (2010) and de Figueiredo (2010).

Within the theoretical and empirical framework summarized by Elhorst (2003), we build a large dataset for 23[1] colombian cities, and categorize the variables into one of the latter groups, as explained in the next section.

3 Methodology and Data

Factor Analysis Methods[2]

Factor analysis methods are multivariate statistical techniques used to handle and process great amounts of information contained in a table (or several tables) in a very exploratory fashion. In order to understand the underlying structure of a table with several observations (usually understood as individuals) and several realizations (variables), the researcher should analyse how "related" or how "different" are individuals or variables. It is also of his interest to assess simplified representations of the data in lower dimensional spaces, which shall allow him to make more straightforward conclusions about trends, unobservable variables and likeness between individuals and between variables. One way to reach this goal is by summarizing the whole structure of the table into a new set of a few non-observable variables, called factors, such that those factors are the variables driving the heterogeneity among observations in the dataset.

Let $\mathbf{X}$ be a matrix with I rows (individuals) and K columns (variables). It is clear that $\mathbf{X}$ belongs to an Euclidean space in $\mathbb{R}^K$ and therefore, we can set a metric in order to measure the distance between any two rows x_i and x_j, $i, j \in I$, namely $d(i, j)$ with $x_i,\ x_j \in \mathbb{R}^K$. This euclidean metric is defined by the $K \times K$ matrix $\mathbf{M}$ and can be different from the canonical metric. For sake of simplicity upon interpretation, $\mathbf{M}$ is usually set as a diagonal matrix in most factor analysis applications[3]. In fact, when $\mathbf{M}$ is diagonal, the distance between points i and j is hence expressed as $d^2(i, j) = \sum_{k \in K}(x_{i,k} - x_{j,k})^2 \cdot m_k$. Since $m_k \in \text{diag}(\mathbf{M})$ weights the influence of each variable $k \in K$ when computing the distance between points i and j, $\mathbf{M}$ is usually understood as the "columns weights" matrix.

The shape of the individuals cloud is completely defined by the coordinates of $\mathbf{X}$ and its associated metric $\mathbf{M}$. However, when calculating the inertia structure (variance) of $\mathbf{X}$, the weight associated to every point $x_i \in X$, p_i, $i \in I$, enters in the computation. These weights are ordered

[1] 8 metropolitan areas that consist of 52 municipalities, and 15 capital cities, as shown below.

[2] For introductory, yet comprehensive, references about multivariate statistical analysis methods see Escofier and Pagès (2008), Peña (2002), and Johnson and Wichern (2007).

[3] When $\mathbf{M}$ is set to be the canonical metric then $\mathbf{M} \equiv I_K$, where I_K is the identity matrix of order K. However, when the metric is not represented by a diagonal matrix, their interpretation is usually more difficult: The scalar product between two rows, x_i and x_j is then $< x_i, x_j >_{\mathbf{M}} = x_i' \mathbf{M} x_j = x_j' \mathbf{M} x_i$.

in a diagonal matrix $\mathbf{D}$ of rank I. Recall that the more heterogeneous the individuals $x_i \in \mathbf{X}$ are, the richer the inertia structure of $\mathbf{X}$ is.

Now let u_h, $h \in H$, be a vector in $\mathbb{R}^K$, and let $F_{u_h} = \mathbf{X}\mathbf{M}u_h$ be the projection of $\mathbf{X}$ over u_h. Note that the variance of $\mathbf{X}$ projected over u_h is $\sum_{i \in I} p_i [F_{u_h}(i)]^2 = F'_{u_h}\mathbf{D}F_u = u'_h\mathbf{M}\mathbf{X}'\mathbf{D}\ \mathbf{X}\mathbf{M}u_h$. Since factor analysis methods aim to build a new set of orthonormal vectors in a lower dimensional space (i.e. $u_h \in \mathbb{R}^H$, $H \leq K$, $\forall\ h \in H$), such that the inertia of $\mathbf{X}$ projected over each one of them is maximized, we are then interested in the unitary vectors u_h that satisfy

$$u_h \in \arg\max_{u_h \in \mathbb{R}^K} \left\{ \text{Inertia}(\mathbf{X}\mathbf{M}u_h) := u'_h\mathbf{M}\mathbf{X}'\mathbf{D}\ \mathbf{X}\mathbf{M}u_h \right\}, \quad \text{s.t. } u'_h u_h = ||u_h|| = 1; \tag{1}$$

When solving the latter maximization problem, the set of orthonormal vectors $u_h,\ h \in H$ that maximize the projected inertia $F'_{u_h}\mathbf{D}F_{u_h}$, are the eigenvectors of the diagonalizable matrix $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}$, ordered according to their associated eigenvalues ranging from the highest (in absolute value), λ_1 , to the lowest, λ_K. Note that, by construction, the inertia projected over u_h will be λ_h, for each $h \in H$. The latter means that $\text{Inertia}(\mathbf{X}\mathbf{M}) = \sum_{k \in K} \lambda_k$.

Principal Component Analysis (PCA), Correspondence Analysis (CA) and Multiple Correspondence Analysis (MCA) are specific cases of this general factor method, and therefore each one has its own specification for matrices $\mathbf{X}$, $\mathbf{D}$ and $\mathbf{M}$. For a detailed presentation of each method see Escofier and Pagès (2008, Chapters 1-4) and Greenacre (2007).

Dealing with mixed datasets

It is clear that several research topics in a wide range of disciplines study statistical units that are simultaneously described by joint sets of *quantitative* (numerical) and *qualitative* (categorical) variables, and *contingency tables*. Although these tables can be separately analysed using PCA, MCA or CA respectively, none of these methods are capable of dealing with a global table constructed as the juxtaposition of several groups of continuous and categorical variables and frequency tables defined on the same set of individuals (rows), like the one shown in Figure 1.

In this $I \times K$ matrix we have J groups of variables, distributed among J_q quantitative groups, J_c categorical groups and J_f frequency tables. Note that $J_q + J_c + J_f = J$, which is the total number of groups in the global table. Also, each group has K_j variables, which means that $\sum_{j \in J} K_j = K$. For notational purposes, x_{ikj} corresponds to a numerical realization (quantitative variable) and z_{ikj} is a dichotomous variable that assigns 1 if x_i belongs to category k in K_j or 0 if not (categorical variable). f_{ikj} is the ratio of the number of occurrences of x_i for variable $k \in K_j$ to the total number of realizations on the contingency table, i.e. $f_{ikj} = x_{ikj} / \sum_i \sum_k x_{ikj}$.

As a response to the issue of working with mixed data, Multiple Factor Analysis (MFA) was developed as a principal axis method that deals with a whole set of quantitative and categorical variables. In MFA the distance between rows is simultaneously determined by both numerical and categorical variables, which is an advantage in comparison to the separate analysis approach (PCA and MCA, respectively). Also, since the contribution of a group with several inertia directions will be greater than a one-dimensional group, the evidence of richer global inertia structure will be summarized by those factors computed through MFA. See Escofier and Pagès (1994, 2008) and Pagès (2002, 2004) for a detailed explanation on MFA.

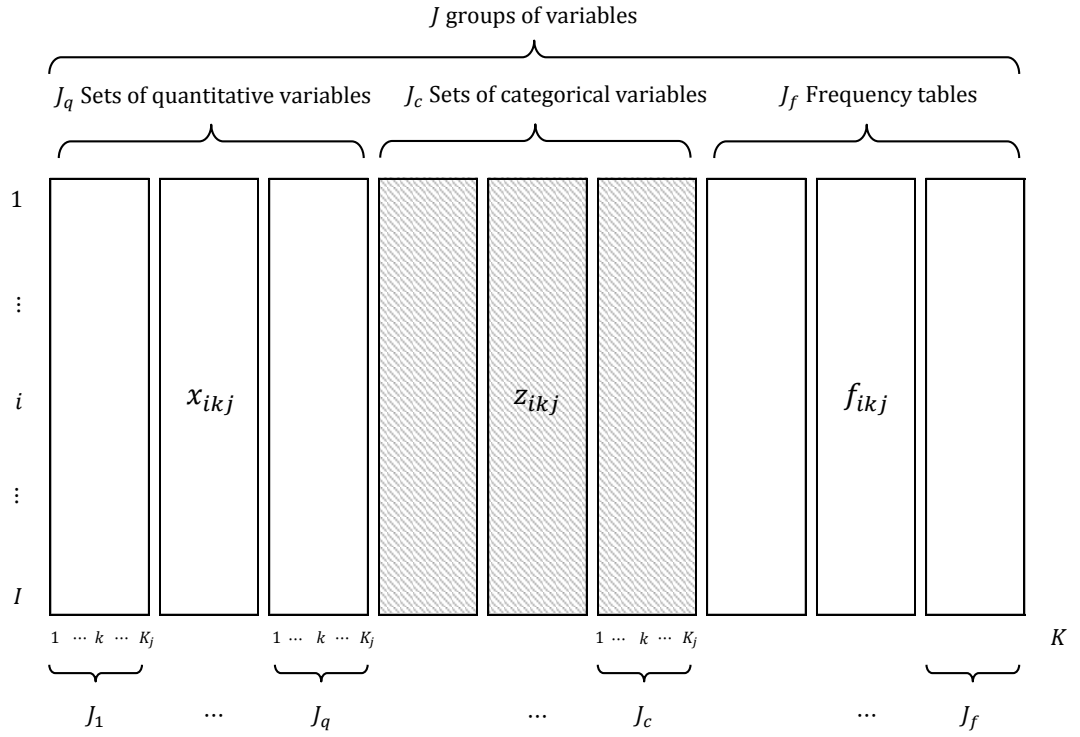


Figure 1: Global table. Adapted from Bécue-Bertaut and Pagès (2008).

MFA balances the weight of each group of variables when computing the inertia of the global table projected onto some vector u_h. This is achieved when the weight of each variable k belonging to a certain group j, m_k^j, is standardized by the first eigenvalue computed in each individual analysis, λ_1^j, i.e. new columns weights are m_k^j/λ_1^j, $\forall\ k \in K$, $\forall\ j \in J$. This procedure allows the researcher to avoid the risk of having a single group dominating the first factor resulting from the global analysis.

More recently Bécue-Bertaut and Pagès (2004, 2008) presented the MFACT, an axial method that extends the MFA to the case in which global tables also contain several contingency tables. In this case, separate analysis results for contingency tables are those obtained when performing CA to the contingency tables in the global table. MFACT can be seen as a general factor method applied to a global table $\mathbf{X}$ subject to some previous transformations (which depend on the nature of the variables), with an specific metric $\mathbf{M}$ and the rows weights $\mathbf{D}$. Matrices are specified in Table 1.

Dealing with a mixture of quantitative, categorical and frequency tables in the global analysis brings some issues when deciding which weights are assigned to individuals. On PCA and MCA (quantitative and categorical tables) individual weights are set according to the user's preferences and are usually fixed to be uniform across all rows ($p_i = 1/I$). However on a multiple contingency table, individual weights are determined by the row margins ($p_i = f_{i\cdot\cdot}$, where $f_{i\cdot\cdot} = \sum_{k\in K}\sum_{j\in J} f_{ikj}$). MFACT can operate under any specification of matrix $\mathbf{D}$. We set $\mathbf{D}$ as in Bécue-Bertaut and Pagès (2008) (i.e. $p_i = f_{i\cdot\cdot}$), to favour cities that have greater populations and to avoid distorted results influenced by uniform individual weights.

In addition, this method supports usual principal axes methods features, such as supplemen-

	Quantitative Variables	Categorical Variables	Frequency Tables
X	$\frac{x_{ikj}-\bar{x}_{kj}}{s_{kj}}$	$\frac{z_{ikj}-(\sum_{i\in I} p_i \cdot z_{ikj})}{\sum_{i\in I} p_i \cdot z_{ikj}}$	$\frac{f_{ikj}-\left(\frac{f_{i\cdot j}}{f_{\cdot\cdot j}}\right)\cdot f_{\cdot kj}}{f_{i\cdot\cdot}\cdot f_{\cdot kj}}$
M	$\frac{1}{\lambda_1^j}$	$\frac{\sum_{i\in I} p_i \cdot z_{ikj}}{\sum_{k\in K_j}\sum_{i\in I} p_i \cdot z_{ikj}\lambda_1^j}$	$\frac{f_{kj}}{\lambda_1^j}$
D	$p_i = f_{i\cdot\cdot} = \sum_{k\in K}\sum_{j\in J} f_{ikj}$		

Table 1: MFACT matrices. Adapted from Bécue-Bertaut and Pagès (2008).

tary projections (both from individuals and variables) and superimposed graphical representations (Escofier and Pagès, 2008).

Data

The dataset used in this paper is a table consisting of 23 rows, one for each city, and 182 variables, grouped among 23 groups. Analysis is restricted to variables for year 2010[4]. Cities considered in this study are: Bogotá and its surrounding municipalities (Soacha, Mosquera, Funza, Madrid, Chía, Cajicá, Cota, La Calera, Tenjo, Tabio, Sibaté, Zipaquirá and Facatitivá), Medellín and its surrounding municipalities (Bello, Barbosa, Copacabana, La Estrella, Girardota, Itagüí, Envigado, Caldas and Sabaneta), Cali and its surrounding municipalities (Palmira, Yumbo, Jamundí, Candelaria, La Cumbre, Vijes and Florida), Barranquilla and its surrounding municipalities (Galapa, Soledad, Puerto Colombia, and Malambo), Bucaramanga and its surrounding municipalities (Floridablanca and Girón), Cúcuta and its surrounding municipalities (Villa del Rosario, Los Patios and El Zulia), Pereira and Dosquebradas, Manizales and its surrounding municipalities (Neira, Chinchiná, Villamaría and Palestina), Pasto, Ibagué, Montería, Cartagena, Villavicencio, Tunja, Florencia, Popayán, Valledupar, Quibdó, Neiva, Riohacha, Santa Marta, Armenia and Sincelejo.

The 182 variables are classified into two categories: *quantitative variables* (119) and *contingency (frequency) tables* (63). Quantitative groups are: demographic variables (*Demo_c*, 5), participation (*Part_c*, 11), inter-regional migration (*Mig_c*, 4), commuting (*Mob_c*, 7), market structure (*Mktst_c*, 21), regional growth (*Regw_c*, 14), market potential (*Mktp_c*, 11), educational attainment (*Edu_c*, 12), wages (*Wag_c*, 12), unionisation (*Unio_c*, 1) and economic and social

[4]Some variables used in this paper display cyclical behaviour along with the business cycle phase. However, these variables are structural determinants of regional unemployment differentials, which means that we expect them to be trend steady or even stationary over a short span of time. That is why we also expect the results not to heavily depend on the year for which this exercise is computed. In addition to the lack of a proper cross-sectional times series data set that allowed us to do comparisons between years, we chose 2010 as our basis year because by 2010 the colombian economy had just overcome the 2008 - 2009 financial crisis and GDP growth rate for that year (4,0%) was close to what has been considered as its potential growth rate.

barriers (*Esbr_c*, 21). In addition, 12 contingency tables that count for 63 variables are constructed: age structure (*Demo_f1*, 4), age structure for men (*Demo_f2*, 4), age structure for women (*Demo_f3*, 4), marital status for men (*Part_f1*, 5), marital status for women (*Part_f2*, 5), waged employment structure (*Mktst_f1*, 2), employment structure by occupational position (*Mktst_f2*, 9), employment structure by economic sector (*Mktst_f3*, 10), educational attainment structure for unemployed population (*Edu_f1*, 5), educational attainment structure for employed population (*Edu_f2*, 5), educational attainment structure for working age population (*Edu_f3*, 5) and educational attainment structure for inactive population (*Edu_f4*, 5). The dataset was constructed with information obtained from several sources including the National Statistics Administrative Department (DANE), Ministry of Education (MEN), Ministry of Finance (MHCP), Department for Social Prosperity (DPS), Economic Commission for Latin America and the Caribbean (ECLAC), Observatory for the Colombian Caribbean (Ocaribe) and the Central Bank of Colombia (Banco de la República). Dataset is available from the authors upon request. Due to the lack of information availability in eight variables for two cities[5], we use the method presented in Husson and Josse (2013) to handle missing data in our sample.

4 Results

Before presenting any results, it is of great importance to recall that endogeneity or causality between variables do not represent any drawback to principal axis computations[6]. These methods reduce dimensionality in order to facilitate analysis and exploratory data conclusions, while keeping maximum the distance between individuals (i.e. differentials or heterogeneity among cities in our sample). The results and interpretations presented here are just of descriptive nature and do not intent to provide any theoretical explanation to labour market dynamics. It is also important to stress the fact that no assumption is made on the multivariate distribution of our data. This means that no probabilistic results shall arise from an MFACT exercise and therefore, we will not make any kind of statistical inference from our data (at least at this stage).

We also point out that we project variables belonging to groups *Demo_f1, Mktst_f1* and *Edu_f3* as supplementary (as seen in Figure 3, groups in italics) so these groups do not add any extra information to the principal axes computations[7].

Interpreting results from factor analysis methods involves various steps. After revising the results for each separate analysis, we find rich variance structure for each group of variables and therefore, we claim that this provides enough evidence for supporting a MFACT approach. Firstly, we assess how many factors serve as a satisfactory representation of the information (inertia) contained in the original dataset. We choose the subset of eigenvectors associated to the eigenvalues that are greater than 1, i.e. factors that in sum explain a high percentage of variance. We find that our sample's inertia ($\sum_{j\in J}\lambda_j = 26,6$) is summarized in a 76,1% by the first five principal axes. Eigenvalues are shown in Figure 2.

[5]Namely: Herfindahl and Hirschman's index for exports diversity, weighted distance to closest markets, Herfindahl and Hirschman's index for market diversity, firm's efficiency index, industrial density, store construction costs, registration costs and sale taxes for Quibó and Florencia.

[6]Actually, as explained in the previous sections, since by construction dimensions are orthogonal, possible feedback effects between one or more variables are not of our concern. As a matter of fact, PCA and other analogous methods are commonly used on regression analysis when explanatory variables are not lineally independent.

[7]Note that by construction *Edu_f1* + *Edu_f2* + *Edu_f4* = *Edu_f3*, *Mktst_f2* + *Mktst_f3* = *Mktst_f1* and *Demo_f2* + *Demo_f3* = *Demo_f1*

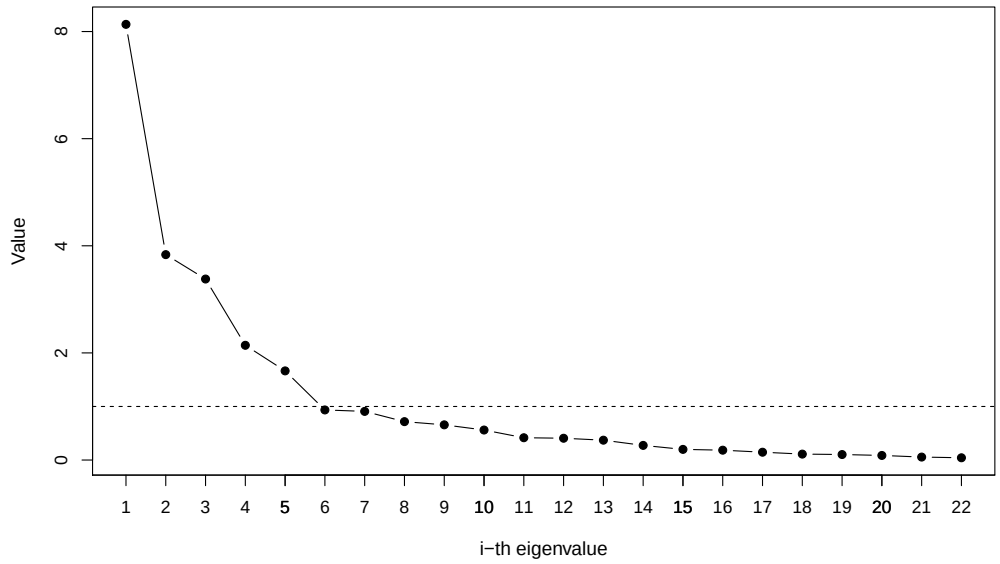


Figure 2: Eigenvalues: Scree plot for Global Factor Analysis.

It is relevant to emphasize the fact that for our analysis, following the approach presented by Bécue-Bertaut and Pagès (2004, 2008), row weights (p_i, $i = 1,2,...,23$) are set according to the aggregate contingency table *row margins*, which is consistent with (and almost identical to) the populations' share of each city on the total national. That is, larger and more inhabited cities are given more importance when computing the gravity centres on the principal axes methods.

Interpreting Principal Axes

We name our chosen factors after the groups of variables (and variables) that contribute the most to the inertia summarized by each dimension and that are also highly correlated with the eigenvectors (see Table 2). We also control for the *quality* of the projections, which is given by the squared cosine of the angle between the point (either group, variable or individual) and the axis (the eigenvector), as explained in Escofier and Pagès (2008).

The first principal axis retained from the global analysis ranks cities depending on the population's educational attainment, workforce productivity and their occupational positions. This axis explains 32.8% of the total variance ($\lambda_1 = 8,7$) and is associated with variables such as number of waged workers, people with 13 or more years of formal education (i.e. those who have completed college or graduate programs) or nominal and real wages. We understand this dimension as an "*index for quality of labour supply*".

The second factor has high loadings on participation variables and educational attainment of unemployed and inactive population, as shown in Figure 3. This dimension counts for almost 16,0% of the total variance ($\lambda_2 = 4,2$). This axis distinguishes cities with high remittances per capita and a demographic structure biased towards elder population (65 years or older) to the negative values side. Cities projected along this side are also characterized by displaying high unemployment rates for low skilled workers. In contrast, projections along the positive side are those of cities that exhibit higher unemployment rates on skilled population and show low participation on the labour market. For these reasons, the second axis has been labelled as the dimension for "*participation and skilled job demand frictions*".

	Contribution (%)					Correlation				
Groups	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Demo_c	3.13	1.69	2.80	**15.22**	3.26	0.61	0.38	0.34	0.57	0.24
Part_c	0.15	**13.88**	5.16	0.26	0.10	0.12	0.77	0.42	0.28	0.06
Mig_c	2.16	0.04	**13.85**	0.99	**11.79**	0.44	0.08	0.69	0.43	0.44
Mob_c	5.95	2.31	1.32	0.08	0.62	0.72	0.31	0.21	0.04	0.10
Mktst_c	6.18	1.13	6.89	3.66	4.39	0.74	0.34	0.51	0.44	0.63
Regw_c	2.39	0.72	3.00	6.08	**11.46**	0.53	0.35	0.33	0.41	0.45
Mktp_c	6.68	0.60	3.00	4.75	5.14	0.76	0.16	0.32	0.32	0.29
Edu_c	5.48	0.83	**17.63**	**18.04**	2.70	0.89	0.26	0.79	0.70	0.42
Wag_c	**7.54**	2.71	0.12	3.91	5.62	0.81	0.37	0.17	0.29	0.31
Unio_c	1.16	0.60	4.31	**15.00**	0.33	0.32	0.16	0.38	0.57	0.07
Esbr_c	6.22	2.67	8.82	4.04	7.90	0.77	0.34	0.65	0.41	0.36
Demo_f2	7.36	9.23	1.11	0.75	0.66	0.82	0.68	0.33	0.23	0.12
Demo_f3	6.69	9.71	0.70	0.34	0.47	0.78	0.67	0.22	0.18	0.12
Part_f1	5.90	4.98	8.65	0.24	7.06	0.73	0.52	0.62	0.14	0.39
Part_f2	6.29	5.53	6.80	0.33	6.13	0.75	0.59	0.50	0.22	0.39
Mkts_f2	**7.97**	1.72	3.60	3.61	7.99	0.84	0.44	0.43	0.47	0.39
Mkts_f3	6.40	3.26	1.71	7.64	**11.69**	0.81	0.55	0.28	0.54	0.53
Edu_f1	1.57	**17.59**	3.25	6.07	2.95	0.43	0.87	0.49	0.58	0.34
Edu_f2	**8.63**	6.03	4.56	4.09	1.01	0.88	0.57	0.68	0.57	0.23
Edu_f4	2.14	**14.74**	2.71	4.88	8.75	0.49	0.80	0.68	0.56	0.40
Suplementary groups										
Demo_f1	6.96	9.60	0.87	0.47	0.54	-	-	-	-	-
Mkts_f1	7.82	0.29	1.39	1.77	5.88	-	-	-	-	-
Edu_f3	6.24	8.95	3.78	4.51	1.27	-	-	-	-	-

Table 2: Inertia contribution and correlations of each group of variables.

Third axis, which explains 12,9% of the total variance ($\lambda_3 = 3,4$), is highly related to education, migration and economic and social barriers groups. Along this dimension, cities are projected according to their public education coverage, specially at middle and high school levels, and their migration profile, which means that this axis has high loadings on net migration rates between and within (from rural to urban spaces) regions. It is also relevant to state that royalties per capita are also highly and negatively correlated with the third axis. This dimension summarizes the differences that arise between cities on an opportunity basis. We label this axis as a "*public education efficiency and migration vulnerability index*".

Fourth axis has also high loadings on education, principally on public education coverage at a basic level. However, it is also closely related to demographic (race) and unionisation variables. Higher values on this dimension mean high percentage of population belonging to an afro descendent ethnic in comparison to the rest of the sample (positive values). This axis also identifies cities with high proportion of unionised workers (negative values of the axis). These results support the idea of thinking about this axis as an "*non-wage rigidities dimension*" on the labour market. This factor counts for 8,1% of the total inertia ($\lambda_4 = 2,2$). Fifth axis explains 6,3% of the total variance ($\lambda_5 = 1,7$), and it has high loadings on migration and regional growth groups, as well as the labour market structure (sectoral employment distribution). This axis is interpreted as the "*economic diversity axis and labour absortion capacity*".

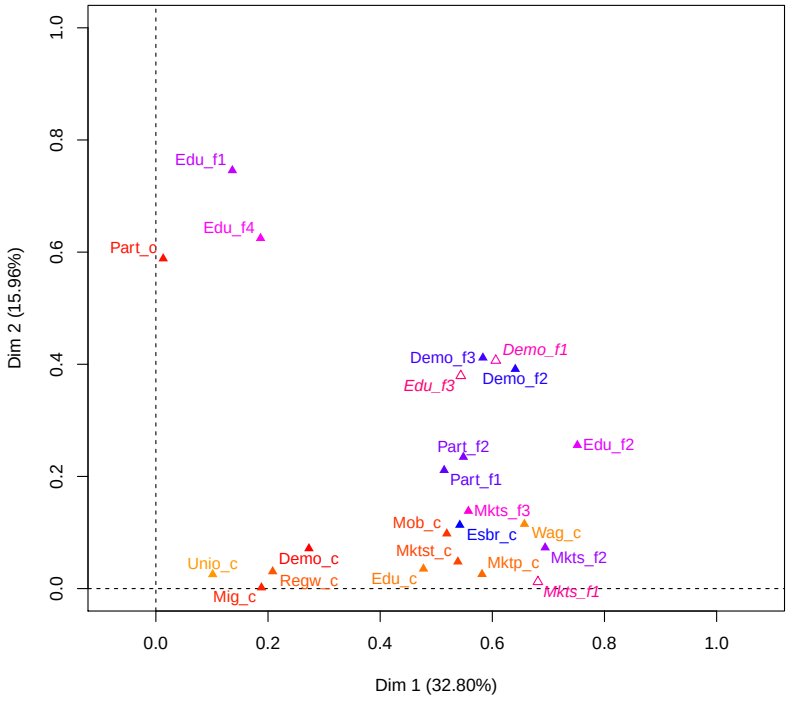


Figure 3: Groups coordinates over the first principal plane.

Interpreting cities' (individuals) projections

We now project the matrix **XM** over the planes formed by pairs of factors. These projections allow us to explore the heterogeneity of a cloud of I individuals in $\mathbb{R}^K$, by analysing the inertia structure of separate clouds of I individuals in $\mathbb{R}^2$. Projections for the first principal plane are just the ordered pairs $(F_{u_1}(i), F_{u_2}(i))$, for $i = 1, ..., 23$. What is most relevant about these projections is that we can assert general conclusions from the fact that interpretations over each principal factor are independent from the others.

FIRST PRINCIPAL PLANE (FIRST AND SECOND DIMENSIONS - FIGURE 4A)

Cities projected along the first dimension are ordered according to their workforce skills, wages and occupational positions. On the positive side of the axis we identify cities with high levels of qualified labour supply, better salaries and more stable, productive and promising job positions, as in Bogotá, Medellín and Bucaramanga. In contrast, cities with low human capital stocks, low wages and poor educational conditions, such as Quibdó, Florencia, and Riohacha, are projected onto the left side of the axis, as seen in Figure 4a. When comparing those cities on an individual basis, we find huge differentials over the variables that contribute the most to the first axis computation.

Pereira, Cali and Manizales are projected onto the negative side of the second dimension, in opposition to cities like Tunja, Cartagena and Quibdó, which are projected onto the positive side. This is consistent with the meaning of this axis: cities where low (high) unemployment rates prevail for educated population are located to the negative (positive) side of the axis, as confirmed by the data on regional educational attainment cohorts. Also, Pereira, Cali and Armenia are among the cities that receive major inflows of remittances from Spain and the United States, in contrast to Tunja, Riohacha and Quibó. Remittances inflows are thought to discourage participation on

urban labour markets.

Projections for the first principal plane are quite interesting and give an initial insight into the structure of urban labour markets in Colombia (Figure 4a). Distance from each projection to the origin is a measure of how average the cities are along these dimensions: the larger this distance, the more different the cities are from the rest of our sample. There are interesting results on this plane, for example those obtained for Tunja, Quibdó and Pereira.

Tunja, which is projected onto the first quadrant, is understood to have highly qualified workforce, along with Bogotá, Medellín and Bucaramanga (positive coordinate along the first axis), but it also displays high unemployment rates for skilled workforce, as in Quibdó, Valledupar and Riohacha. We argue that this might provide evidence of mismatching between labour supply characteristics and labour demand needs. We believe that the local economy is not demanding enough educated workers, since it might be biased towards agricultural activities, as suggested by the information on departmental GDP provided by DANE. Another hypothesis is that although population is skilled, the economic activity did not evolve as rapidly as the educational attainment did, acting as a barrier to the creation of proper job positions for educated people. It is important to recall that we do not control for any variable that measures education quality on the cities in our sample, thus leaving some doubts about the presence of that issue in Tunja.

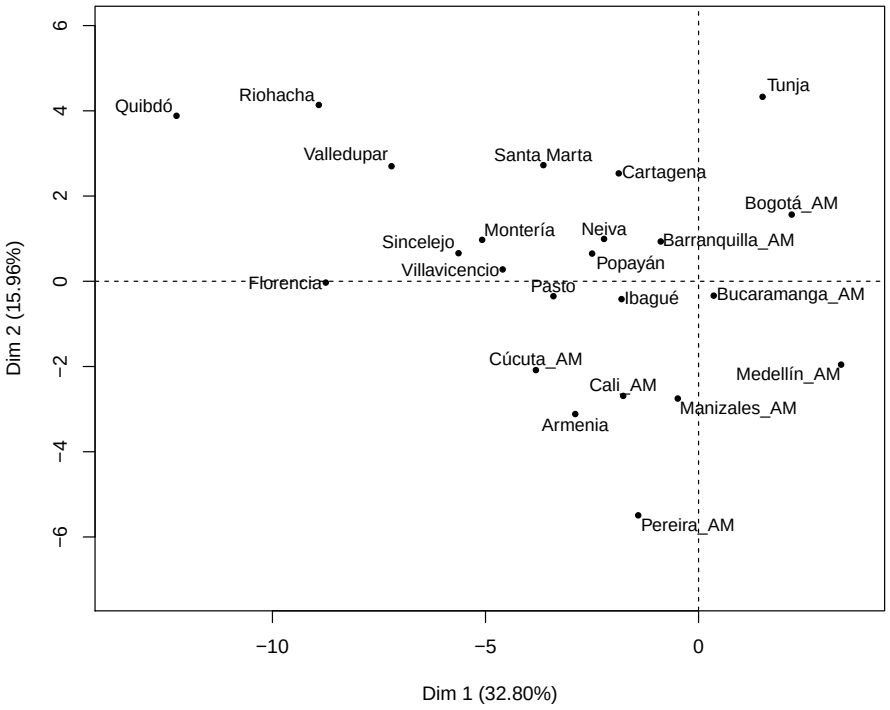
Quibdó, for instance, is a city where economic opportunities appear to be scarce. Poor economic performance and few job positions for skilled people are, among others, factors that determine the lack of willingness of its population to commit themselves the to endeavour of building up better human capital stocks.

Results for Pereira (and some other cities located at the coffee-growers axis region) are also of our special interest. Their projection along the first and second dimensions suggest that these cities are characterized by an old ageing population, high remittances dependence and low skilled workforce. The coexistence of these factors represent difficulties for the accumulation of human capital, since population pyramid is already old and the incentives and facilities to enrol in capacitation programs are not sufficient for the working age population. As a result, these cities have experienced poor economic growth over the past few years, specially in those sectors that are intensive in low skilled labour force (i.e. construction, retail, among others).

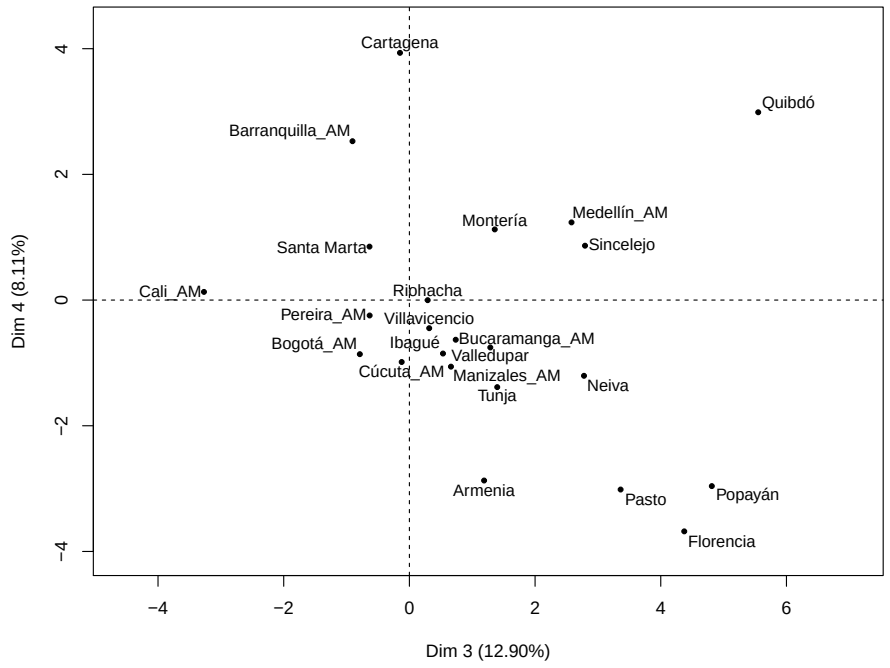
SECOND PRINCIPAL PLANE (THIRD AND FOURTH DIMENSIONS - FIGURE 4B)

Along the third dimension, the most striking differences arise between cities like Quibdó, Popayán or Florencia and Cali or Bogotá, as expected by the information provided by migration variables and public education coverage for those cities. At first glance, the incentives for migration are seemingly higher on the former group of cities: in addition to violence and other political issues (which we do not control for), low wages, low education quality at all scholar stages (despite the efforts on coverage), higher fractions of young population that are not committed to any form of marital relationship and an economic structure that is not absorbing low skilled workers and does not provide enough opportunities to youth, are, among others, the main reasons that encourage migration between and within regions where those cities belong.

Projections along the fourth dimension distinguish cities with a high percentage of unionised workers, such as Popayán, Florencia and Neiva, from those with a relatively low fraction, as in



(a) Individuals representation: First and second axis.



(b) Individuals representation: Third and fourth axis.

Figure 4: Projections over the first and second principal planes.

Barranquilla and Cartagena. This finding suggests that labour markets in cities projected onto the negative side of this axis face rigidities originated in the labour supply side market power. Also, this axis seems to classify cities depending on the average size of households and other participation variables: cities located along the Caribbean region (most of them projected onto the positive side of the axis) are characterized by larger families and very low female participation rates in comparison to other cities in our sample. Finally, there are some demographic characteristics that also contribute to the computation of this axis. Cities in which the afro descendant share of population is higher are projected to the positive side, as in Barranquilla, Quibdó and Cartagena. Another noticeable fact is that departments belonging to the Caribbean region display better economic performance on the energy sector.

Results do acknowledge for differentials on city projections along the first four principal axes. This means that heterogeneity does exist on the determinants of regional unemployment differentials. On a statistical basis, the *inertia ratio* is a criterion that confirms the evidence of labour markets heterogeneity among cities. We first state that according to Huygens theorem, total inertia can be decomposed into *between* and *within* inertia[8]. This measure represents the portion of total inertia that is explained by differences between cities, and not by differences within groups of variables for each city. High values of this ratio for an axis mean heterogeneity between cities along that dimension, while low values might count for diverse results within groups of variables for the same city.

Our results show that the first dimension distinguishes cities differentials better than other axes (*inertia ratio* = 47%), which means that cities are actually heterogeneous along this axis. However, other dimensions are also informative about structural labour market differentials, but in a decreasing fashion, as shown in Table 3.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Ratio	0.47	0.24	0.21	0.15	0.10

Table 3: Inertia ratio.

Results obtained from this methodology provide solid evidence of heterogeneity between colombian cities on variables that are supposed to determine unemployment differentials and labour markets performance, as suggested by Arango (2013). However, one of the most interesting findings in this paper is that cities with high unemployment levels do not necessarily share the same underlining structure on an economic, demographic, educational or even cultural basis. It is clear that there are high disparities between regions on unemployment rates, but not all arise from the same reasons.

5 Clustering

Exploratory Data Analysis (EDA) refers to all statistical procedures that facilitate multivariate data description and visualization. One of the core objectives of EDA is to analyse the resemblances and differences between individuals from a multidimensional point of view (Husson et al.,

[8]This distinction means that if n points in a $\mathbb{R}^K$ space can be grouped into J classes, *within inertia* corresponds to the variance accounted by points belonging to the same group (i.e. inertia of the points with respect to the gravity centre of the group), while *between inertia* is the variance explained by different groups (i.e. inertia between gravity centres of each group). In this particular case, each point corresponds to a partial representation of the individual for each group of variables, while the gravity centre corresponds to the individuals' projection for the global analysis.

2010). Clustering methods are statistical procedures that aggregate individuals (cities in our case) into different groups based on a distance measure and an agglomeration criterion previously defined. Therefore, a natural step when studying the heterogeneity across cities on the structural determinants for regional unemployment differentials, is to apply clustering techniques to our results in order to determine which cities are the most alike and which cities are the most different.

Clustering and multivariate data analysis techniques are complementary methods on EDA, as suggested by Lebart (1994, p.162). Studying the similarities between individuals on a lower dimensional Euclidean space leads to a better understanding of the structure of the data. Since we have already overcome the dimensionality issue by applying the MFACT method to our dataset, we want to group cities that share the same characteristics along the five principal axes we chose in the principal axes analysis step. We follow the approach proposed by Husson et al. (2010), which combines MFACT and both hierarchical (Ward's criterion) and partitional (*k-means* algorithm) clustering methods to robustly describe the differences between groups of individuals. This approach assures that each cluster is fully characterized by the chosen factors and the variables that determine them.

In order to describe and interpret each partition, we measure the association between the cluster (understood as a categorical variable) and each (group of) quantitative variable and each frequency (contingency) set, as described in Bécue-Bertaut and Pagès (2008, pages 3261-3262): classical *t*-tests for differences in means are used to compare whether the cluster's means in quantitative variables are statistically different from those of the whole sample (Lebart et al., 2000), and permutation tests are used to check if frequency variables are over- or under- represented in comparison to the entire set of individuals (Lebart et al., 1998). Variables for which the null hypothesis $H_0 : \mu_{cluster} = \mu_{global}$ is rejected at a certain level of significance, are sought to describe that cluster.

Resulting clusters are then composed of those cities that share the same core characteristics when assessing the structural determinants of regional unemployment differentials, both by groups or individual variables. What is most interesting in this step of the analysis is the fact that, following clustering procedures and evaluating their results based on statistical criteria (hypothesis rejection at 95% and 90% levels of significance), we find that differentials on unemployment rates are associated with different factors across colombian cities.

First cluster: Quibdó, Florencia, Riohacha and Valledupar

Cities belonging to this cluster are Quibdó, Florencia, Valledupar and Riohacha. Although their individual unemployment rates were not the highest among the sample and are not significantly different from those observed for the urban areas (12,4% on average for 2010), there are notorious differences on other variables that determine life quality and human capital formation. Factor analysis results suggest that this cluster is conformed by those cities that are farthest from the gravity centre, positioned to the outer part o the second quadrant on the first principal plane. This means that the first cluster groups those cities that are, on a statistical basis, different in many aspects from the others. Cluster 1 characterizes those cities in which educational attainment is outstandingly low and where poor economic and social perspectives influence on participation and human capital accumulation decisions.

When going deeper into the group's characteristics, we find that cities conforming the first cluster are statistically different on a demographic basis: with more than a 50% share, younger

population (from 0 to 25 years old) has an important participation on the grand total. Also afro descendant and indigenous groups are the most representative ethnics on these cities. Younger people are more likely to be unemployed, mostly due to lack of expertise and education, among other reasons (Furnham, 1985). On the other hand, there is empirical evidence of race discrimination on employment and wage setting (Darity and Mason, 1998), meaning higher unemployment rates for afro descendants and indigenous people.

Migration variables show that these cities are net migrant recipients. This is because these are capital cities of departments where conditions are not favourable for rural population, mostly due to security problems, lack of rural development policies and poor coverage of health and education systems. This feature also impacts other characteristics that determine labour market conditions. Average Educational attainment is very low for these cities: illiteracy rates are the highest among the sample and occupied workforce has the lowest levels of years of schooling, as projections over the first principal axis confirm. In addition, access opportunities to communication services and technology are scarce, as internet service coverage and computer usage indicators reveal. Educational outlook suggests that equilibrium between supply and demand on these labour markets face mismatching frictions. This problem yields high unemployment rates, specially for those who have not provided themselves with enough human capital.

This situation allows low value added and less productive economic activities to have an important participation on their GDP, as concluded from economic structure variables. Over the past decade, mining activities counted for, on average, about a third of their GDP (30%), almost ten times higher than the total national share over the period 2000-2010 (3,4%). Shares of industry, commerce and finance are significantly below the national average (3,2% vs 14,0%; 8,7% vs 12,4% and 6,0% vs 20,7%, respectively). Although commodities prices have risen over the past few years and despite of the weight of mining activities on the cluster, average wages are about 20% lower than those paid in other cities of our sample. This finding confirms the high inequality that is commonly reported for people living on these cities.

In summary, this cluster is conformed by cities where inhabitants have low educational and productive skills, while economic structure is biased towards activities that are not workforce intensive and that are not chained with other sectors that add more aggregated value to the economy. It might be worth to note that this cluster groups cities whose characteristics are more of a consequence than a cause of malfunctions on labour (and other) markets. Poverty and inequality have deeper roots into political and economic issues that are not entirely related to poor labour conditions.

Second cluster: Popayán, Pasto, Montería, Neiva, Villavicencio and Sincelejo

Cities belonging to this cluster are Popayán, Pasto, Montería, Neiva, Villavicencio and Sincelejo. It is important to notice that unemployment rates for these cities are somewhat heterogeneous, but they do still share underlining characteristics about their labour market structure that are statistically significant when labelling these cities into a cluster. As an illustration, Popayán, Pasto and Montería display unemployment rates that are higher that the one reported for the total urban areas (18,1%, 16,0% and 15,1% vs 12,4%), but unemployment rates for Neiva, Villaviencio and Sincelejo are not far or below that threshold (12,7%, 11,9% and 11,1%, respectively); however, the average ratio of non-salaried workers to total workforce is above the total national ratio (68,1% vs 53,9%), as well the average share of self-employed working population (51,7% vs 46,8%).This feature is by far the most relevant characteristic of this cluster: results show that

cities in this group exhibit high levels self-employment and low enrolment on formal firms.

In addition, there is a larger proportion of unionised workers (6,3% vs 3,4% national), which can be thought as friction for the equilibrium reaching mechanism on these labour markets. It has been shown that unionised manufacturing firms tend to expand at a lower speed than the non-unionised ones (Long, 1993; Hirsch, 1997), which might contribute to overall lower economic performance and a lower labour demand expansion over time. It is important to recall that these unionisation levels are low in comparison to other countries in Latin America and around the world (Blanchflower, 2006; Visser, 2006), which can lead to lower effects of this variable on labour markets performance.

Another characteristic that share these cities is that tertiary sector (i.e. commerce, transport and services in general) has gained importance in these economies over the past few years: the average growth rate for the last decade is 6,6%, greater than the average for the national case (4,3% per year), according to the departmental GDP series published by DANE. This performance has been achieved in great part due to the dynamics of the financial services sector: average growth rates are 6,3% and 4,4% respectively.

Other variables suggest that unemployment rates might hide poor labour conditions on these cities. Both average nominal and real incomes are below the national average, even if that difference is not statistically significant. In addition, although not included in the principal axes computations, it is also important to report that underemployment rates (both subjective and objective) are above the average for urban areas: 33,9% and 15,5% vs 30,1% and 12,9%, respectively. This situation allows us to think that cities belonging to this cluster are mostly characterized by dysfunctional formal labour markets where existing working conditions encourage self-employment and informality, and those conditions are not necessarily reflected on the unemployment rate itself.

Third cluster: Barranquilla, Santa Marta and Cartagena

This group is constituted by the most important capital cities along the Colombian Caribbean region: Barranquilla, Santa Marta and Cartagena. This is the cluster with the lowest average unemployment rate in comparison to the urban areas rate (9,9% vs 12,4%). It is also comparatively low on occupation and participation rates (52,3%, 57,8% vs 57,2%, 65,5%, respectively).

Concerning about participation variables, global participation rates for people under 25 years old and for women are outstandingly below the national average for both demographic groups. Among other reasons, this could be related to the average household size, which is the highest among the clusters in our sample (4,1 persons vs 3,7 for the national average). According to the data, it seems that women continue to take over most of parenting responsibilities and domestic tasks at home, supporting the evidence of lower female participation rates.

Regarding education variables, this group displays the largest number of non-public owned institutions per 100,000 inhabitants (43 institutions on average vs 30 for the national average). It is important to note that according to Viloria (2006), higher education coverage has increased over the past few years along the Caribbean region, quality but it is still low when compared to the national average: low scores records, as measured by the results obtained in ECAES test and low R&D investment levels reported for department in this region suggest that coverage efforts have not been companied by quality improvements. This mismatching problem could be the re-

sult of low quality levels on qualified workforce preparation. In fact, average number of years of schooling are for inactive and unemployed are reported to be higher than their respective national averages, which can be associated to the fact that labour supply characteristics do not match labour demand needs.

Data obtained for these cities supports this hypothesis. For example, the average number of scholar years for unemployed population is the highest among the clusters (11,1 vs 10,1 for the national average). In addition, the average number of qualified unemployed people (with college or graduate education) is much higher than the national average, suggesting that skilled employment absorption in this cluster is insufficient and much lower than for other cities in our sample.

Summing up, unemployment and participation rates in this cluster are, on average, the lowest in our sample. Also, young people and women participate less in the labour market than in other cities. This is consistent with the high enrolment rates on educational and housing activities. However, it is important to pay special attention to the quality of both higher education and job placements, since low unemployment rates may be due to the lack of dynamic and inclusive institutions on labour markets, in addition to the fact that overall participation rates are already low.

Fourth cluster: Pereira, Armenia, Manizales, Ibagué, Cúcuta and Cali

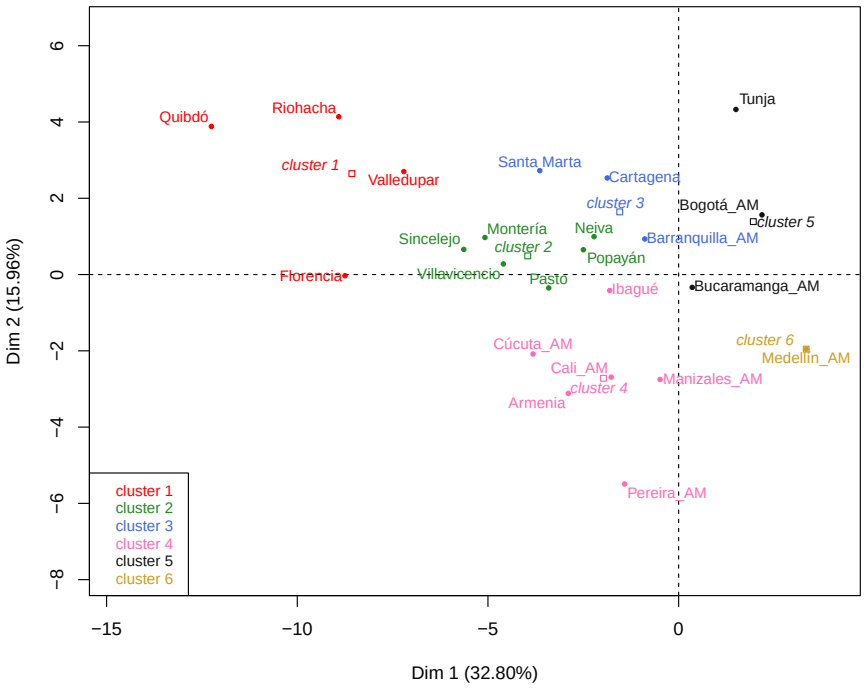
Cities belonging to this cluster are Pereira, Armenia, Manizales, Ibagué, Cúcuta and Cali. This cluster is mainly characterized by their demographic composition - which is skewed towards older population; and their population's educational attainment - which is biased towards population having few years of schooling (less than 10).

These cities exhibit lower gross birth rates than the average of urban areas (15‰ vs 22 ‰), which suggests that the population pyramid tends to reverse faster in this group. This feature results in, among others consequences, a greater proportion of dependent population and lower levels of education in the long run, since educational levels are already low in these cities and, given the progressive population ageing, incentives for human capital formation are decreasing across time.

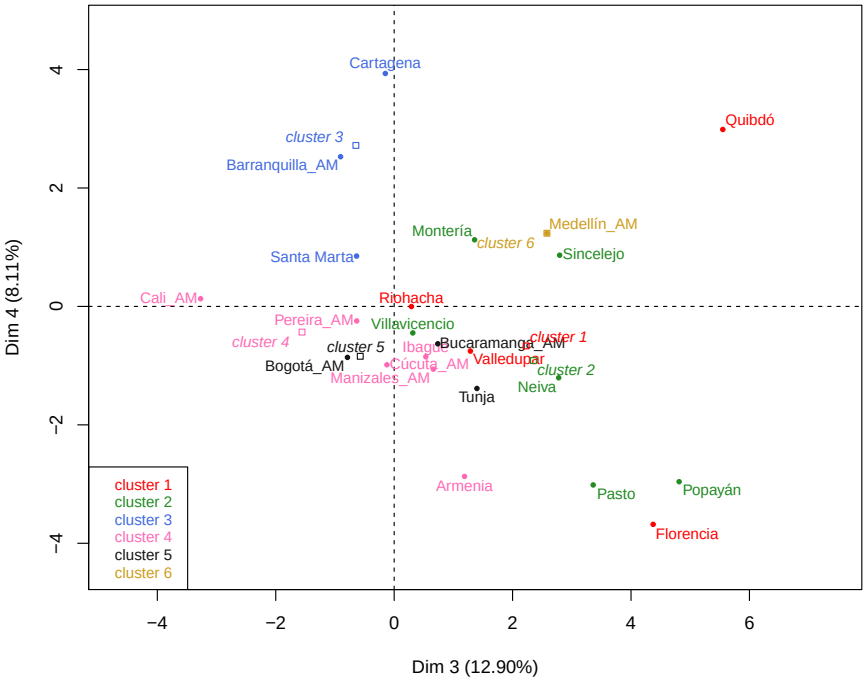
On the other hand, remittances per capita (three times the national average)[9] and high hidden unemployment rates suggest a possible discouragement phenomenon that lowers people's incentives to participate in the labour market[10]. Data reported for these cities support the hypothesis: participation rates for males that are 25 years old or more and for population that is 45 years old or more are below the national average. Actually, unemployment rate for the latter population segment is the highest in the sample (11% compared to the national average of 8%).

[9]This indicator is highly influenced by Pereira, whose per capita remittances received in 2010 amounted to USD $554, followed by Armenia (USD $286), Cali (USD $263), Cucuta (USD $138), Manizales (USD $99), and finally Ibagué with USD $36. Excluding Ibagué (which stands in the eleventh position), cities in this cluster are the largest recipients of remittances.

[10]The average participation rate of this group of cities is 63,2%, barely above the national average. However Cali and Ibagué have higher participation rates than the rest of the cities that conform this cluster. In the case of Ibagué, Lopez (2007) explains that the high rate of participation in this city is caused by low quality of employment, as evidenced on the high underemployment levels and high degree of informality in its labour market. High unemployment rates cannot be attributed uniquely to the global participation rate levels, but to a problem of job quality as well.



(a) City clusters representation: First principal plane.



(b) City clusters representation: Second principal plane.

Figure 5: Cluster Projections over the first and second principal planes.

Another remarkable fact is that living costs (both in levels and annual variations) are, in average, lower than the rest of the economy for the period 2008-2010, as deduced from both the food CPI inflation (1,1% vs 1,9% national average) and total CPI inflation (2,3% vs 2,6%). Low living costs reinforce the hypothesis of lower incentives for people to improve their income levels (by increasing productivity and / or educational level) and to participate in the labour market, since remittances incomes and low living costs make reservation wage increase[11].

Migration variables play an important role in this cluster as well. The average net migration rate is negative and is the second lowest after cluster 1. For the colombian case there is evidence of qualified workforce migration to other cities with better economic performance and more favourable labour conditions, such as Bogotá or even abroad. Concerning this topic, Cepeda (2012) reports a clear concentration of highly skilled human capital in Bogotá, whereas people from other cities in Colombia do not have enough incentives to stay or go back to their own home-towns. This "brain drain"- as known in the international literature - has consequences such as low economic development, low productivity and low wages, which again cause second round effects and create a vicious circle on labour markets performance (Eggert et al., 2010). In fact, the average share of qualified labour force in these cities, both employed and unemployed, is the lowest among the clusters.

On the labour demand side, we find that this cluster has experienced the lowest economic growth on the period 2000-2010 (3,1% on average vs 4,1% for the colombian economy). Weak economic performance is generalized for all sectors, but it is most worrying on the secondary (industry and construction) and tertiary (commerce and services) sectors, which are supposed to be labour intensive sectors.

Interestingly, this cluster groups cities that exhibit the highest unemployment rates in the sample (Pereira, 20,5%; Ibagué, 17,6%; Manizales, 17,6%; and Armenia 16,3%), along with Cúcuta (14,0%) and Cali (13,9%). Our hypothesis is that high unemployment rates in these cities arise due to the coexistence of high non-skilled labour supply levels, low incentives for participation, older population predominance, rigidities on human capital accumulation and an economic structure that is not inclusive for non-qualified available workforce. Results show that the mismatch between supply characteristics and demand needs is definitely a major issue on labour markets in these cities, which can in turn determine long run structural unemployment (Yarce, 2000).

Fifth cluster: Bogotá, Tunja and Bucaramanga

Bogotá, Bucaramanga and Tunja are the cities that belong to this group, which is mainly characterized by both high educational level of the population and higher wages. The average scholar years of the working age population, specifically those of the inactive and employed with more than 15 years of schooling stand out as important characteristics for the fifth cluster. On average 31% of the working age population is qualified (with higher education) in this group of cities[12] , in contrast to the national average of 22%.

[11] Arango et al. (2013) argues that remittances reduce incentives to participate in labour markets by increasing workers' reservation wage and the likelihood of discouragement for unemployed population, by financing long periods of job searching.

[12] Along with Cartagena, the cities of this cluster have the highest shares of qualified the working age population. Tunja stands out with 38%, followed by Bogotá (28%), Bucaramanga and Cartagena (27%).

Cities in this group also display the highest average GDP per capita, and both real and nominal incomes, which are higher by approximately 30% in comparison to the rest of the cities in our sample, only surpassed by Medellín and its metropolitan area (Cluster 6).

In this cluster, the percentage of workers employed on the financial intermediation sector is higher than the national average (2,1% vs 1,3%), as well as those employed on real estate activities (9,1% vs 6,3%). These shares reflect the degree of specialization of these economies on activities oriented towards the provision of services (which weights about 50% of the consumption basket of colombian households). It highlights the industry participation and its good performance during 2000-2010 decade (5,1% on average in this cluster compared to 3,5% nationally).

The average global participation rate in this group of cities is also higher than the national average (67,4% vs 65,5%), mainly caused by increased female participation in comparison to the rest of cities in our sample (61,3% vs 54%). Despite of the higher female labour supply availability, unemployment rate for women is the lowest among all the other clusters (13%), and total unemployment rate is one of the lowest in the country.

In summary, this cluster has very high levels of skilled workforce supply and a higher demand and absorption capacity for this kind of labour than that observed for the rest of the cities in our sample. Results show that there is definitely low mismatching issues on these cities[13], since labour supply responds to demand needs for skilled and productive workforce. This scenario has recently propelled good labour market performance, which in turn allows average unemployment rate for this group to be lower than the reported for the urban areas (11,4% vs 12,4%). However, Gutiérrez (2010) states that the employability in Bogotá has declined over the past few years, so the relatively low unemployment rate has been supported by an increase in informal employment.

Sixth cluster: Medellín

Medellín and its metropolitan area form a cluster by themselves, mainly characterized by market potential variables. These variables are associated with population density, very high industrial density and the lowest average distance to major markets, factors which in principle would yield lower levels of unemployment rate (because of the matching and higher labour demand, as explained in section 2). Also the average household income (both nominal and real) for this cluster is above the national average in about 30%.

However, this city does not exhibit an unemployment rate below the national average (13,9% versus 12,4%). Despite that the industrial sector absorbs much of working population in comparison to the national average (21,2% vs 12,2%), and even though services oriented sectors (which are labour intensive) count for almost 50% of the economic activity, there exists a presumed mismatching issue in this city as well. It is noteworthy that although educational attainment levels are above the national average, demand for qualified workforce seems to be just partially fulfilled: skilled unemployed population share is just 28%, lower than the 30% for our sample and below the 40% in cluster 5. Our hypothesis is that demand is presumably requiring more skilled workforce than labour supply can provide in Medellín. This may be due to the low incentives that many young people have to invest in human capital, given the violent environment in which they live, as argued by Medina et al. (2011).

[13]Except for Tunja, where the unemployment rate is somewhat higher than the urban areas average (12,9% vs 12,4%), despite of the very high levels of skilled workforce in this city.

6 Conclusions

The heterogeneity found on urban labour market indicators in Colombia has not been widely studied, yet no definite conclusions on this topic have been formulated. This paper aims to explore deeply the relations between variables that have been theoretically and empirically assessed to determine the differentials on regional unemployment. Following Elhorst (2003), we study a large dataset in order to stablish similarities or differences between colombian cities based on principal axes methods (MFACT, Bécue-Bertaut and Pagès 2004, 2008), clustering techniques and statistical criteria (Husson et al., 2010). Our results suggest that there is evidence of disparities on structural variables that define the performance of regional labour markets. Particularly, our most relevant result is that cities that display high unemployment rates do not necessarily share the same characteristics, that is, frictions that originate unemployment are not the same across colombian cities.

In order to distinguish which cities do share the same characteristics, clustering techniques are applied to the principal axes results. Groups of cities that are conformed give an important and wide outlook of the colombian labour market structure (Figure 6). For example, we find that high unemployment rates on cluster 4 obey primarily to the mismatch between labour supply and labour demand that results from the lack of educated workforce and the presumed need for qualified workers, and also to low participation incentives due to the presence of high levels of per capita remittances; while unemployment problems on cluster 2 have origins on the high levels of self-employment and the risks associated to this kind of positions. As suggested in the influential work of Overman et al. (2002), having in mind that not all cities or regions share the same structural problems nor react to the same national-based labour institutions allow policy makers to propose and execute better local policies focused towards unemployment reduction and inequalities redressing, and not applying country-based actions that would eventually have no visible effect on regional unemployment rates. Therefore, this type of analysis matters and it is an important tool for achieving national and local governments goals. It is worth noting that in many cases clusters are conformed by nearby cities, which suggests that regional effects are also influenced by geographical positions, as suggested in Overman et al. (2002) and Garcilazo and Spiezia (2007).

In summary, we briefly describe each cluster based on our findings about variables that determine regional labour market differentials: Firstly, *Cluster 1* (Quibdó, Florencia, Valledupar and Riohacha) is conformed by cities where poverty and lack of strong institutional background prevail. As shown in the appendix section, this cluster is statistically different from others on the first, second and fifth principal axes, that is, on labour force quality and demand for skilled workforce due to the prevailing economic structure. *Cluster 2* (Popayán, Pasto, Montería, Neiva, Villavicencio and Sincelejo) is characterized by the great importance of self-employment on the occupational position distribution. Although unemployment rates are not surprisingly high in these cities, this situation leads to high rates of informality, low average income and high underemployment rates, which can influence labour markets performance in the long run.

Cluster 3 (Cartagena, Barranquilla and Santa Marta) is statistically different along the fourth dimension, i.e. malfunctions on labour markets in these cities arise mostly from non-wage rigidities, as evidenced on the outstandingly low participation rates, specially for female working age population. This situation yields low unemployment rates, but average wages and income suggest

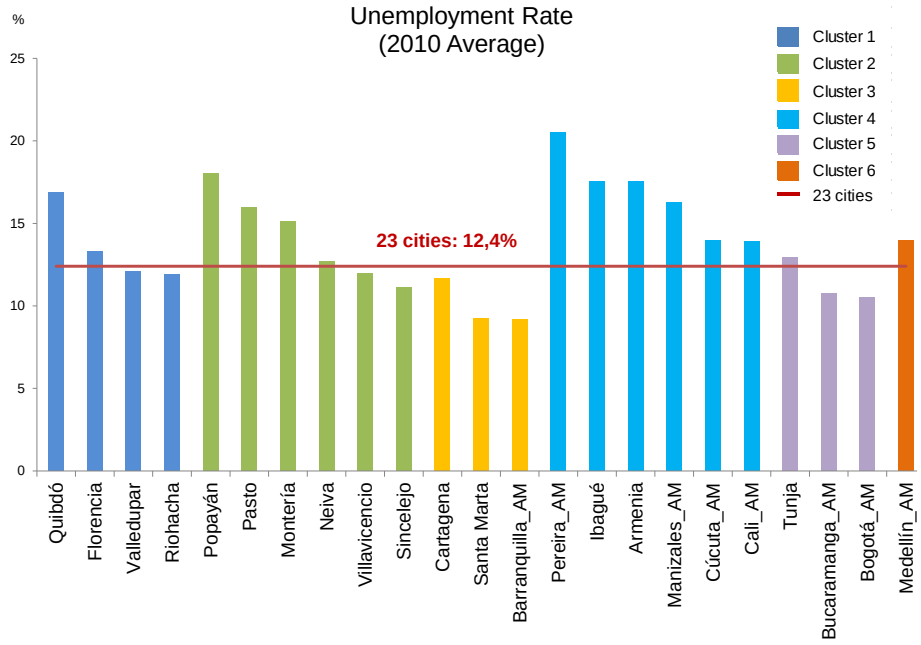


Figure 6: Average unemployment rates for 23 colombian cities (2010).

low quality of job positions.

Cluster 4 is perhaps the most interesting and complex group, since it is conformed by cities that displayed very high unemployment rates in 2010 (Pereira, Ibagué, Armenia, Manizales, Cúcuta and Cali). Statistical tests suggest that this cluster is different from others along the second and third dimensions, i.e. on participation and skilled workforce demand frictions and on migration and opportunities vulnerabilities, as presented in section 4. Low educational attainment and a population pyramid biased towards old population suggest that scholar efforts might not have the expected impact on the average number of scholar years, since older population have less incentives (and abilities) to build up new skills. Cepeda (2012) lists different plausible policies designed to reduce "brain drain"in Colombia, such as local and departmental universities strengthening and scholarship programs for research activities on key economic sectors.

Results for *Cluster 5* (Bogotá, Tunja and Bucaramanga) and *Cluster 6* (Medellín) show that these clusters are conformed by the most prosperous cities in Colombia. However, even though unemployment rates are low, salaries are above the national average and educational attainment is fairly high, there are still some unresolved social and cultural issues on Medellín that influence labour market performance and yield a higher unemployment rate than the average for the metropolitan areas. As pointed out by Medina et al. (2011), military intervention has only short term effects, since the crime culture is deeply embedded in this city. In this case, an alternative solution is a socially oriented intervention with more human investment and inclusive citizen participation on the city's consolidation (Valencia et al., 2008).

Our results provide an useful insight into labour markets' structures in Colombia. However, there are still some minor differences on unemployment rates between cities belonging to the same cluster (as shown in Figure 6) that are not fully captured by differentials on variables used in this paper. We encourage future works to give a deeper insight into each one of these clusters in order

to explore such inner heterogeneities. In addition, since our aim is descriptive and exploratory, we do not focus on giving recommendations about what should or should not be done by policy makers. Instead of that, we just give evidence of regional heterogeneity on variables that explain differentials between urban unemployment rates. Our findings suggest that some cities share common structural characteristics that allow for variety on unemployment rates level in colombian urban areas. However, it is clear that if city-based actions on participation encouragement are taken, or local incentives to low-skilled labour intensive sectors are given away on some cities, or even if regional youth educational programs are supported, unemployment rates will likely decline over time.

References

Albagli, E., García, P., and Restrepo, J. (2004). Labor market rigidities and structural shocks: An open-economy approach for international comparisons. *Central Bank of Chile Working Papers*, 263.

Arango, L. E. (2013). Mercado de trabajo en Colombia: Suma de partes heterogéneas. In Arango, L. E. and Hamann, F., editors, *El mercado de trabajo en Colombia: hechos, tendencias e instituciones*. Banco de la República.

Arango, L. E. and Hamann, F., editors (2013). *El mercado de trabajo en Colombia: hechos, tendencias e instituciones*. Banco de la República.

Arango, L. E., Montenegro, P., and Obando, N. (2013). El desempleo en Pereira: ¿solo cuestión de remesas? In Arango, L. E. and Hamann, F., editors, *El mercado de trabajo en Colombia: hechos, tendencias e instituciones*. Banco de la República.

Bande, R., Fernández, M., and Montuenga, V. (2008). Regional unemployment in Spain: Disparities, business cycle and wage setting. *Labour Economics*, 15:885 – 914.

Barón, J. (2013). Sensibilidad de la oferta de migrantes internos a las condiciones del mercado laboral en las principales ciudades de Colombia. In Arango, L. E. and Hamann, F., editors, *El mercado de trabajo en Colombia: hechos, tendencias e instituciones*. Banco de la República.

Bécue-Bertaut, M. and Pagès, J. (2004). A principal axes method for comparing contingency tables: MFACT. *Computational Statistics & Data Analysis*, 45(3):481 – 503.

Bécue-Bertaut, M. and Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, 52(6):3255 – 3268.

Biffl, G. (1998). The Impact of Demographic Changes on Labor Supply. *Austrian Economic Quarterly*, 4.

Blanchard, O. and Katz, L. (1992). Regional Evolutions. *Brookings Papers on Economic Activity*, 23(1):1 – 75.

Blanchflower, D. (2006). A Cross-country Study of Union Membership. *IZA Discussion Papers*, 2016.

Blundell, R. and MaCurty, T. (1999). Labor Supply: A review of alternative approaches. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3A, chapter 27, pages 1559 – 1695. Amsterdam: Elsevier Science.

Brueckner, J., Thisse, J.-F., and Zenou, Y. (2002). Local labor markets, job matching and urban location. *International Economic Review*, 43(1):155 – 171.

Brunello, G., Lupi, C., and Ordine, P. (2000). Regional disparities and the italian NAIRU. *Oxford Economic Papers*, 52:146 – 177.

Burdett, K. and Mortesen, D. (1998). Wage Differentials, Employer Size and Unemployment. *International Economic Review*, 39(2):257 – 273.

Cahuc, P. and Zylberberg, A. (2004). *Labor Economics*. MIT Press.

Cepeda, L. (2012). Fuga interregional de cerebros - El caso colombiano. *Banco de la República. Documentos de trabajo sobre Economía Regional*, 167.

Chawla, M., Betcherman, G., and Banerji, A. (2007). *From Red to Gray: The "Third Transition"of Aging Populations in Eastern Europe and the former Soviet Union*. The World Bank.

Da Rocha, J. and Fuster, L. (2006). Why are fertility rates and female employment ratios positively correlated across OECD countries? *International Economic Review*, 47(6):1187 – 1222.

Darity, W. and Mason, P. (1998). Evidence on Discrimination in Employment: Codes of color, codes of gender. *The Journal of Economic Perspectives*, 12(2):63 – 90.

de Figueiredo, E. A. (2010). Dynamics of regional unemployment rates in Brazil: Fractional behavior, structural breaks and Markov switching. *Economic Modelling*, 27:900 – 908.

Détang-Dessendre, C. and Gaigné, C. (2009). Unemployment duration, city size and the tightness of the labor market. *Regional Science and Urban Economics*, 39:266 – 276.

Eggert, W., Krieger, T., and Meier, V. (2010). Education, unemployment and migration. *Journal of Public Economics*, (94).

Elhorst, P. (2003). The mystery of regional unemployment differentials: Theoretical and empirical explanations. *Journal of economic surveys*, 17(5):709–748.

Escofier, B. and Pagès, J. (1994). Multiple Factor Analysis (afmult package). *Computational Statistics and Data Analysis*, 18(1):121 – 140.

Escofier, B. and Pagès, J. (2008). *Analyses factorielles simples et multiples; objectifs, méthodes et interprétation.* Dunod.

Farber, H. (1986). The analysis of union behavior. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 2, chapter 18, pages 1039 – 1089. Amsterdam: Elsevier Science.

Fleisher, M. and Rhodes, G. (1976). Unemployment and the Labor force Participation of married men and women: A simultaneus model. *The Review of Economic and Statistics*, 58:398–406.

Freeman, R. (2009). Labor regulations, unions and social protection in developing countries: Market distorsions or efficient institutions? *NBER Working Papers*, (14789).

Furnham, A. (1985). Youth unemployment: a literature review. *Journal of Adolescence*, 8(2):109 – 124.

Galvis, J. (2002). Integración regional de los mercados laborales en colombia: 1984-2000. *Banco de la República. Documentos de trabajo sobre economía regional*, 27.

Gamarra, J. (2005). ¿Se comportan igual las tasas de desempleo de las siete principales ciudades colombianas? *Banco de la República. Documentos de trabajo sobre economía regional*, 55.

Garcilazo, J. and Spiezia, V. (2007). Regional Unemployment Clusters: Neighborhood and State Effects in Europe and North America. *The Review of Regional Studies*, 37(3):282 – 302.

Gomes, F. A. and da Silva, C. G. (2009). Hysteresis versus NAIRU and convergence versus divergence: The behavior of regional unemployment rates in Brazil. *The Quarterly Review of Economics and Finance*, 49:308 – 322.

Greenacre, M. (2007). *Correspondence Analysis in Practice*. Chapman & Hall.

Gupta, K. (1975). Factor Prices, Expectations and Labor Demand. *Econometrica*, 43(4):757 – 770.

Gutiérrez, D. (2010). Evolución del mercado laboral de Bogotá 2001 - 2009: Perspectiva de absorción, y calidad del empleo en bogotá. Technical report, Dirección de Estudios Macro - Secretaría Distrital de Planeación.

Hirsch, B. (1997). Unionization and Economic Performance: Evidence on Productivity, Profits, Investment and Growth. In Milhar, F., editor, *Unions and Right-to-Work Laws*, pages 35–70. The Frazer Institute.

Husson, F. and Josse, J. (2013). Handling missing values in multiple factor analysis. *Food Quality and Preference*, 30:77 – 85.

Husson, F., Josse, J., and Pagès, J. (2010). Principal component methods, hierarchical clustering, partitional clustering: why would we need to choose for visualizing data? *Technical Reports - AgroCampus*.

Izraeli, O. and Murphy, K. (2003). The effect of industrial diversity on state unemployment rate and per capita income. *The annals of Regional Science*, 37:1 – 14.

Jaramillo, C., Romero, C., and Nupia, O. (2000). Integración en el mercado laboral colombiano: 1945 - 1998. *Banco de la República. Borradores de Economía*, 148.

Johnson, R. and Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Pearson - Prentice Hall.

Krugman, P. (1995). *Development, Geography and Economic Theory*. The MIT Press.

Krusell, P. and Rudanko, L. (2013). Unions in a frictional labor market. *NBER Working Papers*, (18128).

Lanzafame, M. (2010). The nature of regional unemployment in Italy. *Empirical Economics*, 39:877 – 895.

Lebart, L. (1994). Complementary use of correspondence analysis and cluster analysis. In Greenacre, M. and Blasius, J., editors, *Correspondence Analysis in the Social Sciences*, pages 162–178. Academic Press.

Lebart, L., Morineau, A., and Piron, M. (2000). *Statistique exploratoire multidimensionelle*. Dunod.

Lebart, L., Salem, A., and Berry, L. (1998). *Exploring Textual Data*. Kluwer Academic Press.

Lerman, R. and Schmidt, S. (1999). An Overview of Economic, Social and Demographic Trends Affecting the U.S Labor Market: Final Report. Technical report, The Urban Institute - United States Department of Labor.

Lewis, H. G. (1986). Union relative wage effects. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 2, chapter 20, pages 1139 – 1181. Amsterdam: Elsevier Science.

Long, R. (1993). The Effect of Unionization on Employment Growth of Canadian Companies. *Industrial and Labor Relations Review*, 46(4):691 – 703.

Lottman, F. (2012). Explaining regional unemployment differences in Germany: a spatial panel data analysis. *SFB 649 Discussion Papers*, 26.

Malizia, E. and Ke, S. (1993). The influence of economic diversity on unemployment and stability. *Journal of Regional Science*, 33(2):221 – 235.

Martínez, C. (2013). Descenso de la fecundidad, participación laboral de la mujer y reducción de la pobreza en Colombia, 1990 - 2010. *Serie de Estudios a Profundidad ENDS, 1990 - 2010*.

Medina, C., Posso, C., and Tamayo, J. (2011). Costos de la violencia urbana y políticas públicas: Algunas lecciones de Medellín. *Borradores de Economía*, 614.

Mincer, J. (1981). Union effects: Wages, turnover and job traning. *NBER Working Papers*, (808).

Mincer, J. (1991). Education and unemployment. *NBER Working Papers*, (3838).

Moretti, E. (2010). Local multipliers. *American Economic Review, Papers and Proceedings*, 2(100):1 – 7.

Moretti, E. (2012). *The new geography of jobs*. Houghton Mifflin Harcourt.

Oberst, C. and Oelgemollër, J. (2013). Economic growth and regional labor market development in german regions: Okun's law in a spatial context. *FCN Working Papers*, (5).

Okun, A. (1962). Potential GDP: Its measure and significance. *Cowles Fundation Papers*, (190).

Overman, H., Puga, D., and Hylke, V. (2002). Unemployment Clusters across Europe's Regions and Countries. *Economic Policy*, 17(34):115 – 147.

Pagès, J. (2002). Analyse factorielle multiple appliqué aux variables qualitatives et aux données mixtes. *Revue de Statistique Appliquée*, 50(4):5 – 37.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw Hill.

Pissarides, C. and Wadsworth, J. (1989). Unemployment and the Inter-Regional Mobility of Labour. *The Economic Journal*, 99(397):739 – 755.

Tran, H. P. (2011). *Industrial diversity and economic performance: A spatial analysis*. PhD thesis, University of Nebraska.

Urrutia, M., editor (2001). *Empleo y economía*. Banco de la República.

Valencia, G., Aguirre, M., and Flórez, J. (2008). Capital social, desarrollo y políticas públicas en Medellín, 2004 - 2007. *Estudios Políticos*, 32(1):53 – 83.

Viloria, J. (2006). Educación superior en el Caribe colombiano: Análisis de cobertura y calidad. *Banco de la República. Documentos de trabajo sobre economía regional*, 69.

Visser, J. (2006). Union membership statistics in 24 countries. *Monthly Labor Review*, pages 38 – 49.

Walden, M. (2012). Explaining differences in state unemployment rates during the great recession. *The Journal of Regional Analysis and Policy*, 42(3):251 – 257.

Winters, J. (2013). Human capital externalities and employment differences across metropolitan areas of the USA. *Journal of Economic Geography*, 13:799 – 822.

Yarce, W. (2000). El desempleo estructural y la tasa natural de desempleo: Algunas consideraciones teóricas y su estado actual en Colombia. *Lecturas de Economía*, 52.

Appendices

Group	Variable	Kind of variable	Description	Source
Demographic Structure	TCPP	q	Population Growth Rate (Average 1986-2010)	2005 census estimates (DANE)
	POBT_0-25	f	Population less than 25 years old (2010 estimate)	
	POBT_25-45	f	Population between 26 and 45 years old (2010 estimate)	
	POBT_45-65	f	Population between 46 and 65 years old (2010 estimate)	
	POBT_65+	f	Population more than 65 years old (2010 estimate)	
	POBM_0-25	f	Male population less than 25 years old (2010 estimate)	
	POBM_25-45	f	Male population between 26 and 45 years old (2010 estimate)	
	POBM_45-65	f	Male population between 46 and 65 years old (2010 estimate)	
	POBM_65+	f	Male population more than 65 years old (2010 estimate)	
	POBF_0-25	f	Female population less than 25 years old (2010 estimate)	
	POBF_25-45	f	Female population between 26 and 45 years old (2010 estimate)	
	POBF_45-65	f	Female population between 46 and 65 years old (2010 estimate)	
	POBF_65+	f	Female population more than 65 years old (2010 estimate)	
	PPIN	q	% Indigenous population (2010 estimate)	
	PPAF	q	% Black population (2010 estimate)	
	TMI	q	Infant mortality rate (2010)	
	TBN	q	Gross birth rate (2010)	
Participation	TGP	q	Labor participation rate (2010)	GEIH (DANE)
	TGPM	q	Male labor participation rate (2010)	
	TGPF	q	Female labor participation rate (2010)	
	TDO	q	Hidden unemployment rate (2010)	
	TGPHJ	q	Male labor participation rate less than 25 years (2010)	
	TGPH	q	Male labor participation rate more than 25 years (2010)	
	TGPMJ	q	Female labor participation rate less than 25 years (2010)	
	TGPM	q	Female labor participation rate more than 25 years (2010)	
	EC_H_UL	f	Male - Marital Status: Free union (more than 10 years)	
	EC_H_C	f	Male - Marital Status: Married (more than 10 years)	
	EC_H_SE	f	Male - Marital Status: Divorced (more than 10 years)	
	EC_H_V	f	Male - Marital Status: Widow (more than 10 years)	
	EC_H_SOL	f	Male - Marital Status: Single (more than 10 years)	
	EC_M_UL	f	Female - Marital Status: Free union (more than 10 years)	
	EC_M_C	f	Female - Marital Status: Married (more than 10 years)	
	EC_M_SE	f	Female - Marital Status: Divorced (more than 10 years)	
	EC_M_V	f	Female - Marital Status: Widow (more than 10 years)	
	EC_M_SOL	f	Female - Marital Status: Single (more than 10 years)	
	TH	q	Average household size	
	EPN	q	Women's median age on the fist birth	Profamilia ENDS
	REM	q	Remittances per capita	Banco de la República
Migration	TMN	q	Net migration rate (2010)	DANE
	PREC	q	People received (per 100.000 inhabitants)	DPS
	NPDEP	q	Displacement Register (Arrivals-people) (per 100.000 inhabitants)	
	NHDEP	q	Displacement Register (Arrivals-households) (per 100.000 inhabitants)	
Commuting	ICVSP	q	Public transport use Index (2010)	Authors' calcuations with information from the Urban transporta survey (DANE)
	PSSP	q	Number of passengers carried in Public Service (Buses,less taxi) (per 100.000 inhabitants) 2010	Urban transporta survey (DANE)
	PVSP	q	Number of Public service vehicles in service (2010) (per 100.000 inhabitants)	
	IPCVIV	q	Housing CPI (average 2008-2010)	DANE
	IPCTR	q	Transport services CPI (average 2008-2010)	
	INFVIV	q	Housing CPI change (average 2008-2010)	
	INFTR	q	Transport services CPI change (average 2008-2010)	
Wages	INP	q	Average Nominal Income	GEIH (DANE)
	INM	q	Median Nominal Income	
	IRP	q	Average Real Income	
	IRM	q	Median Real Income	
	SNPA	q	Average nominal wage (wage-salary workers)	
	SNMA	q	Median nominal wage (wage-salary workers)	
	SRPA	q	Average real wage (wage-salary workers)	
	SRMA	q	Median real wage (wage-salary workers)	
	IPCALI	q	Food CPI (average 2008-2010) - Living cost	DANE
	INFALI	q	Food CPI change (average 2008-2010) - Living cost	
	IPCTot	q	Total CPI (average 2008-2010) - Living cost	
	Infla IPCTot	q	Total CPI change (average 2008-2010) - Living cost	

Figure 7: Dataset description.

Regional Growth	PIBT	q	Total GDP Growth (Average 2000-2011) - Departmental	Departmental GDP series - Authors' own calculations (DANE)
	PIB1	q	Primary GDP Growth (Average 2000-2011) - Departmental	
	PIB2	q	Secundary GDP Growth (Average 2000-2011) - Departmental	
	PIB3	q	Terciary GDP Growth (Average 2000-2011) - Departmental	
	PIB_agr	q	Agricultural GDP Growth (Average 2000-2011) - Departmental	
	PIB_min	q	Mining GDP Growth (Average 2000-2011) - Departmental	
	PIB_inds	q	Industry GDP Growth (Average 2000-2011) - Departmental	
	PIB_elec	q	Electricity GDP Growth (Average 2000-2011) - Departmental	
	PIB_cons	q	Construction GDP Growth (Average 2000-2011) - Departmental	
	PIB_com	q	Comercio GDP Growth (Average 2000-2011) - Departmental	
	PIB_trans	q	Transport GDP Growth (Average 2000-2011) - Departmental	
	PIB_finan	q	Financial GDP Growth (Average 2000-2011) - Departmental	
	PIB_serv	q	Services GDP Growth (Average 2000-2011) - Departmental	
	PIB_impu	q	Taxes GDP Growth (Average 2000-2011) - Departmental	
Market Potential	DDM	q	Demographic density (2005)	2005 census estimates (DANE)
	ESML	q	Deviation of unemployment rate from his long-term trend (tight labor market)	GEIH, Author's own calculations (DANE)
	SEMD	q	Average unemployment duration	GEIH (DANE)
	AC	q	Trade Openness	Author's own calculations (DANE)
	DIST_MERC	q	Weighted distance to major markets (Bogotá, Medellín, Cali and Barranquilla)	Ocaribe - SID
	IHHM	q	Herfindahl - Hirschman Market Index	
	EPE_2009	q	Efficiency of Business Processes	
	IDI_2009	q	Industrial Density Index	
	CostConst	q	Construction (cost of building a warehouse)	
	Aemp	q	Starting a business (costs and requirements to form and register a new company)	
	TIT	q	Total Taxes rate 2013	
Economic Structure	IHHP	q	Herfindahl - Hirschman Product Index	Ocaribe - SID
	PART_agr	q	Agricultural GDP Share (average 2000-2011)	Departmental GDP series - Authors' own calculations (DANE)
	PART_min	q	Mining GDP Share (average 2000-2011)	
	PART_inds	q	Industry GDP Share (average 2000-2011)	
	PART_elec	q	Electricity GDP Share (average 2000-2011)	
	PART_cons	q	Construction GDP Share (average 2000-2011)	
	PART_com	q	Trade GDP Share (average 2000-2011)	
	PART_trans	q	Transport GDP Share (average 2000-2011)	
	PART_finan	q	Financial GDP Share (average 2000-2011)	
	PART_serv	q	Services GDP Share (average 2000-2011)	
	PART_impu	q	Taxes GDP Share (average 2000-2011)	
	VAR_PART_agr	q	Change on agricultural GDP share (average 2000-2011)	
	VAR_PART_min	q	Change on mining GDP share (average 2000-2011)	
	VAR_PART_inds	q	Change on industry GDP share (average 2000-2011)	
	VAR_PART_elec	q	Change on electricity GDP share (average 2000-2011)	
	VAR_PART_cons	q	Change on construction GDP share (average 2000-2011)	
	VAR_PART_com	q	Change on trade GDP share (average 2000-2011)	
	VAR_PART_trans	q	Change on Transport GDP share (average 2000-2011)	
	VAR_PART_finan	q	Change on financial GDP share (average 2000-2011)	
	VAR_PART_serv	q	Change on services GDP share (average 2000-2011)	
	VAR_PART_impu	q	Change on taxes GDP share (average 2000-2011)	
	AS	f	Number of wage-salary workers (2010)	GEIH (DANE)
	NAS	f	Number of unpaid wage workers (2010)	
	EMPA	f	Number of private employees	
	EMGOB	f	Number of government employees	
	EMDOM	f	Number of domestic employees	
	EMCP	f	Number of self employed	
	EMPT	f	Number of bosses	
	EMFSR	f	Number of unpaid family workers	
	EMSR	f	Number of unpaid workers	
	EMJ	f	Number of workmen	
	EMO	f	Number of other workers	
	NPAI	f	Number of people working on Real estate activities	
	NPAG	f	Number of people working on Agriculture, fishing, hunting and forestry	
	NPCM	f	Number of people working on Trade, hotels and restaurants	
	NPCN	f	Number of people working on Construction	
	NPEG	f	Number of people working on Gas and Water Supply	
	NPFN	f	Number of people working on financial intermediation	
	NPIN	f	Number of people working on manufacturing	
	NPMN	f	Number of people working on Mines and Quarries exploitation	
	NPSS	f	Number of people working on Services	
	NPTN	f	Number of people working on Transport, storage and communications	

Figure 8: Dataset description (*continued*).

Group	Variable	Kind of variable	Description	Source
Economic and Social Barriers	ICV	q	Life Quality Index (2005)	DNP
	DHV	q	Percentage of households without housing (2005)	2005 census estimates (DANE)
	DHVC	q	Percentage of households without housing (2005) - quantitative	
	NBI	q	Unsatisfied basic needs (2010)	
	DHVQ	q	Percentage of households without housing (2005) - qualitative	
	PIBPC	q	GDP per capita (2010)	Departmental GDP series - Authors' own calculations (DANE)
	PIBPC_SMIN	q	GDP per capita whitout mining (2010)	
	C_PIBPC	q	GDP per capita Growth (average 2001 - 2010)	
	C_PIBPC_SMIN	q	GDP per capita Growth whitout mining (average 2001 - 2010)	
	NH_V	q	Number of households per housing	GEIH (DANE)
	COB_ELEC	q	Percentage of housing without Electric Energy	
	COB_ACU	q	Percentage of housing without Aqueduct	
	COB_SAN	q	Percentage of housing without sewage system	
	COB_BAS	q	Percentage of housing without garbage disposal	
	COB_GAS	q	Percentage of housing without natural Gas	
	EJ_FBKF	q	Gross fixed capital formation per capita Execution (2010)	MHCP, Authors' own calculations
	REGA	q	Income from royalties per capita (2010)	
	TRANS	q	National transfers income per capita (2010)	
	NATH	q	Number of requests for Humanitarian Assistance (2010) per 100.000 inhabitants	DPS
	NNICBF	q	Number of children served by family Welfare Colombian Institute (2010) - per 100.000 inhabitants	
	HOM	q	Murders per 100.000 inhabitants (2010)	Ocaribe - SID
Education	COBP	q	Primary education Coverage (2010)	MEN
	COBS	q	Secundary education Coverage (2010)	
	COBM	q	Vocational media Coverage (2010)	
	IENO	q	Non-public Institutions per 100.000 Inhabitants (2010)	
	IEO	q	Public Institutions per 100.000 Inhabitants (2010)	
	POBA	q	Percentage of Literate population (2010)	
	POBAN	q	Percentage of illiterate population (2010)	
	HOG_SININT	q	Percentage of household without Internet (2010)	GEIH (DANE)
	HOG_SINPC	q	Percentage of household without PC (2010)	
	APED_PET	q	Average education years - Working age population (2010)	
	APED_OC	q	Average education years - Employed population (2010)	
	APED_DE	q	Average education years - Unemployed population (2010)	
	APED_DE_0-5	f	Percentage of Unemployed between 0 and 5 years of education (2010)	
	APED_DE_6-11	f	Percentage of Unemployed between 5 and 11 years of education (2010)	
	APED_DE_12-13	f	Percentage of Unemployed between 12 and 13 years of education (2010)	
	APED_DE_14-15	f	Percentage of Unemployed between 14 and 15 years of education (2010)	
	APED_DE_15+	f	Percentage of Unemployed more than 15 years of education (2010)	
	APED_OC_0-5	f	Percentage of Employed between 0 and 5 years of education (2010)	
	APED_OC_6-11	f	Percentage of Employed between 6 and 11 years of education (2010)	
	APED_OC_12-13	f	Percentage of Employed between 12 and 13 years of education (2010)	
	APED_OC_14-15	f	Percentage of Employed between 14 and 15 years of education (2010)	
	APED_OC_15+	f	Percentage of Employed more than 15 years of education (2010)	
	APED_PET_0-5	f	Percentage of Working age population between 0 and 5 years of education (2010)	
	APED_PET_6-11	f	Percentage of Working age population between 6 and 11 years of education (2010)	
	APED_PET_12-13	f	Percentage of Working age population between 12 and 13 years of education (2010)	
	APED_PET_14-15	f	Percentage of Working age population between 14 and 15 years of education (2010)	
	APED_PET_15+	f	Percentage of Working age population more than 15 years of education (2010)	
	APED_INA_0-5	f	Percentage of Inactive population between 0 and 5 years of education (2010)	
	APED_INA_6-11	f	Percentage of Inactive population between 6 and 11 years of education (2010)	
	APED_INA_12-13	f	Percentage of Inactive population between 12 and 13 years of education (2010)	
	APED_INA_14-15	f	Percentage of Inactive population between 14 and 15 years of education (2010)	
	APED_INA_15+	f	Percentage of Inactive population more than 15 years of education (2010)	
Unionisation	PSIN	q	Percentage of employees unionised	GEIH (DANE)

Figure 9: Dataset description (*continued*).

CLUSTER 1: Quibdó, Florencia, Riohacha and Valledupar

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Commuting	PSSP	(2.89)	2,943.80	11,346.68	1,490.02	6,248.35	0.00
	ICVSP	(2.81)	2.62	5.63	0.71	2.30	0.00
	PVSP	(2.06)	95.07	165.05	47.22	73.03	0.04
Demographic Structure	POBM_0-25	3.69	0.26	0.06	0.11	0.12	0.00
	POBT_0-25	3.68	0.28	0.07	0.11	0.12	0.00
	POBF_0-25	3.66	0.30	0.08	0.11	0.13	0.00
	POBM_45-65	(3.61)	(0.31)	(0.07)	0.10	0.14	0.00
	POBT_45-65	(3.51)	(0.31)	(0.07)	0.10	0.14	0.00
	POBF_45-65	(3.41)	(0.30)	(0.07)	0.10	0.14	0.00
	POBM_25-45	(2.99)	(0.17)	(0.05)	0.10	0.08	0.00
	POBF_65+	(2.88)	(0.32)	(0.06)	0.09	0.20	0.00
	POBT_25-45	(2.84)	(0.15)	(0.05)	0.10	0.08	0.00
	POBT_65+	(2.80)	(0.28)	(0.04)	0.09	0.19	0.01
	POBF_25-45	(2.61)	(0.13)	(0.04)	0.10	0.07	0.01
	POBM_65+	(2.56)	(0.23)	(0.01)	0.10	0.18	0.01
	TMI	2.55	23.39	15.97	10.28	6.26	0.01
	PPAF	2.39	34.78	12.16	36.07	20.39	0.02
	PPIN	2.27	7.73	2.48	7.85	4.97	0.02
	TBN	1.89	28.53	22.39	5.21	7.00	0.06
Economic and social barriers	COB_ELEC	4.43	6.66	1.57	1.73	2.47	0.00
	COB_BAS	3.96	29.27	8.49	11.91	11.31	0.00
	COB_ACU	3.36	38.49	9.79	29.61	18.37	0.00
	ICV	(3.28)	72.84	82.18	5.72	6.13	0.00
	NBI	3.04	49.54	24.57	24.45	17.67	0.00
	COB_SAN	3.04	48.13	17.11	25.72	21.96	0.00
	DHV	2.81	60.24	32.49	24.44	21.23	0.00
	NATH	2.69	4,210.48	1,817.50	2,258.65	1,911.71	0.01
	DHVQ	2.63	43.06	19.91	23.72	18.94	0.01
	PIBPC_SMIN	(2.15)	3,687.14	6,598.59	972.26	2,920.28	0.03
	DHVC	1.86	17.17	12.57	9.37	5.33	0.06
	NH_V	(1.79)	1.02	1.05	0.00	0.03	0.07
	COB_GAS	1.79	60.46	37.26	28.13	27.96	0.07
Economic Structure	NPAG	3.84	4.28	0.85	1.96	1.93	0.00
	NPAI	(3.23)	(0.59)	(0.24)	0.13	0.23	0.00
	NPFN	(2.99)	(0.65)	(0.30)	0.12	0.26	0.00
	NPIN	(2.63)	(0.57)	(0.23)	0.14	0.27	0.01
	PART_inds	(2.55)	0.03	0.11	0.01	0.07	0.01
	PART_min	2.52	0.26	0.09	0.21	0.14	0.01
	PART_finan	(2.50)	0.06	0.13	0.02	0.06	0.01
	NPCM	(2.48)	(0.09)	0.03	0.14	0.11	0.01
	NPMN	2.34	7.28	1.24	11.44	5.56	0.02
	PART_com	(1.98)	0.09	0.12	0.03	0.03	0.05
	PART_impu	(1.94)	0.04	0.07	0.01	0.04	0.05
	PART_trans	(1.90)	0.06	0.07	0.01	0.01	0.06
	PART_serv	1.81	0.25	0.19	0.12	0.07	0.07
	VAR_PART_serv	1.72	0.02	0.01	0.02	0.02	0.09

Figure 10: Clusters description - variables.

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Education	APED_OC_05	3.33	0.53	0.17	0.13	0.23	0.00
	POBAN	3.33	9.09	5.00	1.73	2.64	0.00
	POBA	(3.33)	90.91	95.00	1.73	2.64	0.00
	APED_OC	(3.25)	8.27	9.45	0.40	0.78	0.00
	HOG_SINPC	3.15	82.12	69.71	2.36	8.48	0.00
	HOG_SININT_	3.02	87.23	77.08	2.52	7.24	0.00
	APED_PET	(2.84)	7.91	8.80	0.42	0.68	0.00
	APED_PET_0-5	2.78	0.36	0.12	0.12	0.19	0.01
	APED_OC_14-15	(1.93)	(0.38)	(0.17)	0.10	0.23	0.05
	APED_OC_6-11	(1.90)	(0.11)	(0.02)	0.05	0.10	0.06
	COBP	1.87	125.70	112.58	19.35	15.12	0.06
	IENO	(1.77)	21.82	30.40	7.53	10.45	0.08
	APED_PET_14-15	(1.71)	(0.33)	(0.14)	0.13	0.24	0.09
	APED_PET_06-11	(1.65)	(0.08)	(0.02)	0.02	0.08	0.10
Market Potencial	DIST_MERC	2.97	872.05	572.71	115.76	216.87	0.00
	IHHM	2.92	0.74	0.40	0.15	0.25	0.00
	AC	1.78	0.39	0.21	0.40	0.21	0.08
	CostConst	(1.76)	103.05	164.67	15.15	75.25	0.08
Migration	NPDEP	3.48	2,279.64	923.80	905.76	839.18	0.00
	NHDEP	3.20	583.31	252.04	264.89	222.75	0.00
	PREC	1.95	742.00	341.14	772.94	441.83	0.05
Participation	EC_H_C	(3.61)	(0.49)	(0.09)	0.15	0.24	0.00
	EC_M_UL	3.55	0.57	0.10	0.10	0.28	0.00
	EC_M_C	(3.42)	(0.44)	(0.07)	0.21	0.23	0.00
	EC_H_UL	3.14	0.41	0.07	0.15	0.23	0.00
	EPN	(2.24)	20.25	21.22	0.43	0.93	0.03
	EC_H_V	(2.08)	(0.34)	(0.11)	0.14	0.24	0.04
	EC_H_SOL	2.05	0.09	0.03	0.10	0.06	0.04
	TGPH	1.94	89.17	87.07	2.02	2.33	0.05
	TH	1.94	4.00	3.68	0.20	0.35	0.05
	TH	1.94	4.00	3.68	0.20	0.35	0.05
	EC_M_SE	(1.89)	(0.17)	(0.01)	0.25	0.19	0.06
	TGPF	(1.76)	49.01	54.00	4.84	6.10	0.08
Regional Growth	PIB_trans	2.52	0.08	0.06	0.01	0.02	0.01
	PIB2	2.23	0.11	0.06	0.06	0.05	0.03
Wages	IRM	(1.99)	438,885.80	490,376.70	36,451.37	55,715.90	0.05
	INM	(1.98)	455,375.00	510,704.80	41,727.52	60,094.62	0.05
	INP	(1.96)	630,513.30	744,640.40	45,382.11	125,326.30	0.05
	IRP	(1.95)	607,762.10	714,990.90	37,995.50	118,137.30	0.05

Figure 11: Clusters description - variables (*continued*).

CLUSTER 2 : Popayán, Pasto, Montería, Neiva, Villavicencio and Sincelejo

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Demographic Structure	TBN	2.06	27.57	22.39	4.70	7.00	0.04
Economic and social barriers	TRANS	2.19	57,242.21	24,302.80	60,244.94	41,920.43	0.03
	REGA	2.01	621,129.50	519,777.00	136,425.60	140,255.50	0.04
Economic Structure	EMPA	(2.37)	(0.48)	(0.25)	0.26	0.27	0.02
	EMO	2.26	0.85	0.20	0.98	0.81	0.02
	NAS	2.22	0.32	0.16	0.20	0.20	0.03
	AS	(2.22)	(0.37)	(0.19)	0.23	0.23	0.03
	IHHP	(2.09)	0.21	0.31	0.08	0.13	0.04
	EMCP	1.88	0.34	0.18	0.16	0.24	0.06
	VAR_PART_finan	(1.85)	(0.01)	0.00	0.02	0.01	0.06
	NPCM	1.85	0.11	0.03	0.05	0.11	0.06
Market potential	EPE	(2.09)	0.21	0.31	0.08	0.13	0.04
Participation	EPN	(1.65)	20.67	21.22	0.75	0.93	0.10
Regional Growth	PIB_finan	2.84	0.06	0.04	0.02	0.02	0.00
	PIB3	1.97	0.05	0.04	0.00	0.01	0.05
Unionisation	PSIN	2.10	6.78	4.53	4.30	2.99	0.04

CLUSTER 3: Barranquilla, Santa Marta and Cartagena

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Commuting	PSSP	2.40	19,606.07	11,346.68	4,114.01	6,248.35	0.02
	IPCVIV	(2.14)	102.65	103.91	0.48	1.06	0.03
	ICVSP	2.04	8.22	5.63	1.76	2.30	0.04
	INFVIV	(1.88)	2.74	3.68	0.48	0.91	0.06
Economic Structure	VAR_PART_elec	2.68	0.01	0.00	0.00	0.00	0.01
	NPTN	2.41	0.33	0.08	0.11	0.19	0.02
	EMPT	(2.38)	(0.51)	(0.02)	0.10	0.37	0.02
	VAR_PART_impu	(2.17)	(0.01)	(0.01)	0.01	0.01	0.03
	PART_trans	1.82	0.09	0.07	0.01	0.01	0.07
	PART_impu	1.81	0.11	0.07	0.05	0.04	0.07
Education	APED_INA_6-11	3.20	0.08	(0.02)	0.04	0.06	0.00
	APED_DE_0-5	(2.57)	(0.35)	0.12	0.03	0.33	0.01
	APED_INA_05	(2.44)	(0.19)	0.04	0.08	0.17	0.01
	APED_DE_14-15	2.36	0.35	(0.07)	0.47	0.33	0.02
	APED_PET_06-11	2.32	0.08	(0.02)	0.04	0.08	0.02
	APED_DE	2.30	11.09	10.08	0.03	0.80	0.02
	IENO	2.27	43.45	30.40	7.00	10.45	0.02
	APED_PET_0-5	(2.01)	(0.09)	0.12	0.05	0.19	0.04
	APED_OC_6-11	1.73	0.07	(0.02)	0.06	0.10	0.08
	APED_OC_0-5	(1.67)	(0.05)	0.17	0.05	0.23	0.10
Market potential	IHHM	(1.96)	0.13	0.40	0.07	0.25	0.05
	TIT_DB	1.82	74.77	71.32	1.95	3.43	0.07

Figure 12: Clusters description - variables (*continued*).

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Participation	TGPMJ	(2.44)	24.49	34.95	3.03	7.79	0.01
	TH	2.24	4.12	3.68	0.06	0.35	0.03
	TH	2.24	4.12	3.68	0.06	0.35	0.03
	TGPHJ	(2.13)	36.25	43.89	2.13	6.52	0.03
	TGP	(2.05)	58.22	62.90	1.54	4.15	0.04
	TGPF	(1.81)	47.94	54.00	1.73	6.10	0.07
	TGPM	(1.76)	69.30	72.30	1.59	3.10	0.08
	TGPM	(1.69)	56.05	61.52	3.74	5.88	0.09
Unionisation	PSIN	(1.74)	1.66	4.53	1.28	2.99	0.08

CLUSTER 4: Pereira, Armenia, Manizales, Ibagué, Cúcuta and Cali

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Commuting	IPCVIV	1.83	104.61	103.91	0.71	1.06	0.07
	INFVIV	1.66	4.22	3.68	0.59	0.91	0.10
Demographic Structure	POBM_65+	3.06	0.18	(0.01)	0.12	0.18	0.00
	POBT_65+	2.96	0.16	(0.04)	0.12	0.19	0.00
	POBF_65+	2.84	0.14	(0.06)	0.12	0.20	0.00
	TBN	(2.53)	16.01	22.39	2.53	7.00	0.01
	POBF_45-65	2.32	0.05	(0.07)	0.09	0.14	0.02
	POBT_45-65	2.32	0.05	(0.07)	0.08	0.14	0.02
	POBM_45-65	2.30	0.04	(0.07)	0.08	0.14	0.02
	TCPP_8610	(1.72)	0.02	0.02	0.00	0.01	0.09
	POBF_0-25	(1.69)	(0.00)	0.08	0.06	0.13	0.09
Economic and social barriers	EJ_FBKF	(3.10)	535,907.30	706,998.00	51,878.47	153,921.50	0.00
	REGA	(2.49)	394,464.50	519,777.00	43,075.06	140,255.50	0.01
	DHV	(1.92)	17.84	32.49	5.94	21.23	0.05
	DHVQ	(1.69)	8.46	19.91	4.21	18.94	0.09
	NATH	(1.67)	671.36	1,817.50	396.60	1,911.71	0.09
	DHVC	(1.67)	9.39	12.57	2.36	5.33	0.10
Economic Structure	EMCP	(2.14)	(0.01)	0.18	0.14	0.24	0.03
	NPTN	(2.06)	(0.06)	0.08	0.08	0.19	0.04
	EMPA	1.75	(0.08)	(0.25)	0.14	0.27	0.08
	NPIN	1.74	(0.06)	(0.23)	0.17	0.27	0.08
	PART_finan	1.74	0.17	0.13	0.05	0.06	0.08
	NPCN	(1.70)	(0.03)	0.06	0.08	0.15	0.09
	AS	1.66	(0.05)	(0.19)	0.11	0.23	0.10
	NAS	(1.66)	0.05	0.16	0.09	0.20	0.10
Education	APED_DE_0-5	2.90	0.46	0.12	0.29	0.33	0.00
	APED_DE	(2.78)	9.29	10.08	0.46	0.80	0.01
	APED_DE_12-13	(2.67)	(0.36)	(0.06)	0.23	0.32	0.01
	APED_DE_15+	(2.61)	(0.39)	0.14	0.27	0.56	0.01
	APED_INA_12-13	(2.11)	(0.22)	0.07	0.17	0.38	0.03
	APED_INA_0-5	2.09	0.16	0.04	0.10	0.17	0.04
	APED_DE_14-15	(1.98)	(0.31)	(0.07)	0.23	0.33	0.05
	APED_DE_6-11	1.82	0.04	(0.04)	0.10	0.12	0.07
Market Potential	DIST_MERC	(2.51)	377.53	572.71	128.75	216.87	0.01
	ESML	1.81	0.59	0.19	0.66	0.62	0.07
Migration	NPDEP	(1.85)	367.05	923.80	209.36	839.18	0.06
	NHDEP	(1.80)	107.97	252.04	60.93	222.75	0.07

Figure 13: Clusters description - variables (*continued*).

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Participation	REM	3.28	229.29	82.47	169.63	124.91	0.00
	EC_M_SE	2.78	0.18	(0.01)	0.11	0.19	0.01
	EC_H_V	2.27	0.09	(0.11)	0.16	0.24	0.02
	TGPH	(2.08)	85.33	87.07	1.89	2.33	0.04
	TDO	1.97	0.99	0.77	0.10	0.30	0.05
	TGPMJ	1.95	40.41	34.95	6.56	7.79	0.05
	TH	(1.94)	3.44	3.68	0.16	0.35	0.05
	TH	(1.94)	3.44	3.68	0.16	0.35	0.05
Regional Growth	PIB3	(2.82)	0.03	0.04	0.01	0.01	0.00
	PIB_impu	(2.26)	0.03	0.05	0.02	0.02	0.02
	PIB_trans	(2.07)	0.05	0.06	0.02	0.02	0.04
	PIB_finan	(2.04)	0.03	0.04	0.01	0.02	0.04
	PIBT	(1.97)	0.03	0.04	0.01	0.02	0.05
	PIB_com	(1.77)	0.03	0.03	0.01	0.01	0.08
	PIB_serv	(1.69)	0.03	0.04	0.01	0.01	0.09
Wages	INFALI	(2.31)	1.11	1.87	0.93	0.92	0.02
	IPCALI	(2.18)	100.33	101.18	1.29	1.10	0.03
	SNPA	(1.70)	850,908.50	917,089.70	83,139.36	108,373.80	0.09
	SRPA	(1.65)	819,454.40	881,018.50	86,336.11	103,821.10	0.10

CLUSTER 5: Bogotá, Tunja and Bucaramanga

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Demographic Structure	POBT_25-45	1.81	0.03	(0.05)	0.01	0.08	0.07
	POBF_25-45	1.79	0.03	(0.04)	0.02	0.07	0.07
	POBM_25-45	1.78	0.03	(0.05)	0.01	0.08	0.07
Economic and social barriers	PIBPC_SMIN	3.27	11,861.67	6,598.59	2,674.19	2,920.28	0.00
	HOM	(2.72)	12.07	34.28	6.39	14.86	0.01
	PIBPC	2.43	12,410.48	7,572.41	2,441.31	3,618.87	0.02
	NH_V	2.18	1.09	1.05	0.05	0.03	0.03
Economic Structure	NPFN	2.96	0.12	(0.30)	0.27	0.26	0.00
	IHHP	2.26	0.46	0.31	0.09	0.13	0.02
	PART_impu	1.98	0.11	0.07	0.05	0.04	0.05
	EMCP	(1.93)	(0.08)	0.18	0.12	0.24	0.05
	EMPT	1.93	0.38	(0.02)	0.39	0.37	0.05
	AS	1.90	0.05	(0.19)	0.14	0.23	0.06
	NAS	(1.90)	(0.04)	0.16	0.12	0.20	0.06
	NPAI	1.88	0.00	(0.24)	0.22	0.23	0.06
	PART_inds	1.76	0.17	0.11	0.05	0.07	0.08
	EMPA	1.65	(0.01)	(0.25)	0.16	0.27	0.10
Participation	TGPF	2.19	61.34	54.00	4.10	6.10	0.03
	EC_H_C	2.16	0.20	(0.09)	0.15	0.24	0.03
	EC_M_C	2.06	0.19	(0.07)	0.15	0.23	0.04
	TDO	(1.97)	0.45	0.77	0.10	0.30	0.05
	TGP	1.95	67.36	62.90	4.15	4.15	0.05
	TGPM	1.94	67.79	61.52	1.31	5.88	0.05
	EC_M_UL	(1.77)	(0.17)	0.10	0.16	0.28	0.08
	EC_H_UL	(1.72)	(0.15)	0.07	0.15	0.23	0.09

Figure 14: Clusters description - variables (*continued*).

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Education	APED_PET_15+	3.30	0.44	(0.08)	0.44	0.29	0.00
	APED_INA_15+	3.18	0.50	(0.15)	0.48	0.37	0.00
	APED_OC_15+	2.94	0.38	(0.05)	0.46	0.26	0.00
	APED_PET	2.72	9.81	8.80	0.51	0.68	0.01
	APED_OC	2.49	10.53	9.45	0.60	0.78	0.01
	APED_DE_15+	2.34	0.86	0.14	0.91	0.56	0.02
	APED_INA_14-15	2.32	0.51	(0.00)	0.55	0.40	0.02
	APED_PET_14-15	2.16	0.15	(0.14)	0.04	0.24	0.03
	APED_PET_12-13	2.11	0.24	(0.06)	0.18	0.26	0.04
	APED_PET_0-5	(2.10)	(0.10)	0.12	0.12	0.19	0.04
	COBP	(2.07)	95.38	112.58	5.91	15.12	0.04
	HOG_SINPC	(2.03)	60.24	69.71	3.62	8.48	0.04
	APED_DE_6-11	(1.98)	(0.17)	(0.04)	0.13	0.12	0.05
	APED_INA_12-13	1.97	0.48	0.07	0.58	0.38	0.05
	IEO	(1.96)	7.85	12.96	1.56	4.73	0.05
	APED_DE	1.91	10.92	10.08	0.61	0.80	0.06
	HOG_SININT	(1.91)	69.49	77.08	5.85	7.24	0.06
	APED_INA_0-5	(1.87)	(0.13)	0.04	0.16	0.17	0.06
	POBA	1.81	97.64	95.00	0.26	2.64	0.07
	POBAN	(1.81)	2.36	5.00	0.26	2.64	0.07
	APED_OC_0-5	(1.79)	(0.06)	0.17	0.13	0.23	0.07
	APED_OC_12-13	1.76	0.15	(0.10)	0.06	0.25	0.08
	APED_PET_6-11	(1.75)	(0.09)	(0.02)	0.07	0.08	0.08
Market potential	EPE	2.26	0.46	0.31	0.09	0.13	0.02
	IDI	2.05	0.96	0.34	0.64	0.55	0.04
Regional Growth	PIB_serv	(1.88)	0.03	0.04	0.00	0.01	0.06
	PIB_cons	(1.79)	0.08	0.11	0.02	0.03	0.07
	PIB_com	1.67	0.05	0.03	0.01	0.01	0.10
Wages	INP	3.17	963,325.80	744,640.40	54,865.14	125,326.30	0.00
	IRP	3.17	920,912.00	714,990.90	57,908.35	118,137.30	0.00
	IRM	3.11	585,776.10	490,376.70	10,444.52	55,715.90	0.00
	INM	3.09	612,953.70	510,704.80	9,370.69	60,094.62	0.00
	SNMA	2.25	679,722.20	602,615.30	29,299.96	62,261.42	0.02
	SRMA	2.17	649,659.40	578,854.20	30,454.71	59,197.92	0.03
	SNPA	2.11	1,042,902.00	917,089.70	54,509.05	108,373.80	0.03
	SRPA	2.03	997,070.70	881,018.50	59,799.49	103,821.10	0.04
	IPCALI	1.68	102.20	101.18	0.84	1.10	0.09

CLUSTER 6: Medellín

Group	Variable	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Demographic Structure	POBF_45-65	1.70	0.17	(0.07)	NA	0.14	0.09
	POBT_45-65	1.69	0.17	(0.07)	NA	0.14	0.09
	POBM_45-65	1.67	0.16	(0.07)	NA	0.14	0.10
Economic and social barriers	EJ_FBKF	2.29	1,058,917.00	706,998.00	NA	153,921.50	0.02
Economic Structure	NPIN	1.89	0.28	(0.23)	NA	0.27	0.06
Education	HOG_SININT	(2.01)	62.53	77.08	NA	7.24	0.04
	HOG_SINPC	(1.82)	54.30	69.71	NA	8.48	0.07
Market Potential	DDM	3.91	31,809.83	3,754.71	NA	7,176.46	0.00
	IDI	2.40	1.67	0.34	NA	0.55	0.02
	DIST_MERC	(1.69)	206.12	572.71	NA	216.87	0.09
Participation	EC_M_SOL	1.68	0.13	(0.03)	NA	0.10	0.09
	EC_H_UL	(1.65)	(0.32)	0.07	NA	0.23	0.10
Wages	INP	1.81	971,272.10	744,640.40	NA	125,326.30	0.07
	IRP	1.78	925,559.40	714,990.90	NA	118,137.30	0.07

Figure 15: Clusters description - variables (*continued*).

CLUSTER 1

Dimension	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Dim.5	3.71	4.24	0.00	1.59	2.46	0.00
Dim.2	2.07	2.67	0.31	1.65	2.46	0.04
Dim.3	1.78	2.88	1.20	2.16	2.03	0.08
Dim.1	(3.65)	(9.28)	(3.12)	1.84	3.63	0.00

CLUSTER 2

Dimension	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Dim.3	1.88	2.57	1.20	1.43	2.03	0.06

CLUSTER 3

Dimension	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Dim.4	2.67	2.44	(0.32)	1.26	1.87	0.01

CLUSTER 4

Dimension	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Dim.3	(2.02)	(0.27)	1.20	1.46	2.03	0.04
Dim.2	(3.47)	(2.76)	0.31	1.50	2.46	0.00

CLUSTER 5

Dimension	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Dim.1	2.24	1.35	(3.12)	0.75	3.63	0.03

CLUSTER 6

Dimension	Test Value	Mean in category	Overall mean	Standard Deviation in category	Overall Estándar Deviation	P - value
Dim.1	1.78	3.35	(3.12)	NA	3.63	0.07

Figure 16: Clusters description - dimensions.