Ovidius University Annals
ECONOMIC SCIENCES SERIES
Volume X, Issue 1
Year 2010

OVIDIUS UNIVERSITY PRESS

ISSN 1582 – 9383

OVIDIUS UNIVERSITY ANNALS

ECONOMIC SCIENCES SERIES

Volume X, Issue 1

2010

Portuguese Regional Unemployment Patterns: A k-Means Cluster Analysis Approach

NUNES Alcina
BARROS Elisa
Instituto Politécnico de Bragança
alcina@ipb.pt
ebarros@ipb.pt

Abstract

The k-means cluster analysis technique is an important ally in the study of economic patterns in a multivariate framework. Aware of its analytical importance this paper adopts such method of study to identify groups of Portuguese administrative regions that share similar patterns regarding the characteristics of unemployed registered individuals. The regional distribution of the unemployed individual characteristics is of core importance for the development of public policies directed to fight the unemployment phenomenon, especially in times of crisis. Preliminary results show a clear division of the territory into four regions – north and south and urban and rural areas - that stresses the importance of designing well-directed public labour policies.

Keywords: regional unemployment, labour market, regional dissimilarities.

JEL Classification: J64

1. Introduction

A study of the regional dissimilarities in a labour market can not be limited by a simple descriptive analysis of the associated phenomena. It should establish standards for spatial comparison of the territories subject to analysis in order to develop public policies of both central and regional scope. These policies should be the most appropriate to fight the associated unemployment problems. The generation of employment public policies, used to fight the persistent phenomenon of unemployment, have deserved a particular attention in the Portuguese economy over the last decades, however, little is known about the profile of the registered unemployed individuals along the national administrative regions. That knowledge is of crucial importance to develop public labour market policies specifically targeted to the regional unemployment profiles.

This type of analysis begins to be developed in some economies since it is believed to be of core importance for understanding the phenomenon of unemployment at the aggregate level. Moreover, according to authors like Marelli [4] economies that observe substantial changes in its GDP (a fast growth of the GDP levels, for example) do it at the expense of some regions over others. The consequence is the promotion of the observed regional disparities that could be observed, namely, throughout different profiles of regional unemployment. The so-called European Union enlargement countries had experienced this phenomenon which led to the production of various studies on the phenomenon of regional unemployment [1, 2 and 5].

In this research work, it will be aimed to bound regions of the country - using the district as the territorial unit - according to a specific profile of registered unemployment. So, it will be applied a classification analysis where the territorial units are grouped into classes, according to their similarities observed through the set of explanatory variables presented. The aim is to detect the presence of homogeneity among different districts based on a multivariate statistical method – the k-means clusters analysis methodology – for the year 2009. To reach the mentioned goal it will be taken into account a set of variables made public by the institution that administrates the registers of unemployed individuals in the Portuguese economy – the *Instituto de Emprego e Formação Profissiona (IEFP).*

The article is presented as follows. The next section presents a description of the registered unemployment rate in Portugal taking into

account a set of characteristics that describe the unemployed individuals registered in the national employment services. In Section 3, it is explained, briefly, the methodology selected for the empirical analysis presented in Section 4. This section presents and describes the results obtained there. Section 5 presents the average cluster unemployment profiles and section 6 concludes.

2. Registered unemployment data

The unemployed individuals registered in the Portuguese public employment services of the *IEFP* present a given set of distinctive characteristics that make each of them different from the others. Such features are related with the gender, age, formal education, unemployment spell and relation to a first or new employment. The data concerning these characteristics are openly available in a monthly period base. The data concerning the month of December collects information about the stock of registered unemployed individuals at the end of the respective year.

Table 1. Description of the registered unemployed individuals in Portugal

Characteristics		2009
Gender	Male	236.791
	Female	267.984
Age	<25 years	64.116
	25-34 years	119.441
	35-54 years	229.054
	>=55 years	92.164
Education	Less than 1º CB	27.408
	1º CB	142.665
	2º CB	96.529
	3º CB	99.976
	Secondary	94.442
	Superior	43.755
Position relating labour market	First Employment	37.556
	New Employment	467.219
Unemployment duration	Less than 1 year	329.358
	More than 1 year	175.417
Total		**504.775**

The registered values, here applied, were collected in December of each one of the referred years and represent the stock of register unemployed individuals in the public employment services during the year that did not found a formal employment or other solution to their lack of job problem.

3. Brief introduction to k-means cluster analysis

The seminal work of Tryon [7] introduced the cluster analysis. Such methodology is composed by a set of multivariate statistical methods that include different classification and optimization algorithms which intend to organize information concerning multiple variables and shape homogeneous groups.

In other words the *cluster* analysis develops tools and methods that, given a data matrix containing multivariate measurements on a large number of individuals (or objects), aim to build up some natural groups. The groups or *clusters* should be as homogeneous as possible and the differences among the various groups as large as possible. The *cluster* analysis does not make conjectures about the number of groups or its structures - the groups are based on the similarities among the groups characterized by different ways of calculating the "distance".

Being the adopted variables quantitative variables, the application of the Euclidian Distance method is advised [3]. The distance is defined as the square root of the sum of the squared differences between the values of i and j for all the selected variables $(v = 1, 2, ..., p)$:

$$dij = \sqrt{\sum_{v=1}^{p} \left(X_{iv} - X_{jv} \right)^2}$$

Besides the settlement of the distance among observations, computation method is still necessary to settle the computational method to calculate the distance among groups. Non-hierarchical methods choose in advance the k number of groups which will comprise all the observations. Then all the observations could be divided by the predefined k groups and the best partition of the n observations will be the one that optimizes the chosen criteria. One of the processes that could be applied is the k-means interactive partition method. The method follows the next steps: starts by dividing an initial partition of individuals by the number of clusters previously defined; computes for each cluster the respective centroid and the distances between each individual and the centroid of the various groups; transfers each individual to the cluster relatively to which presents the minor distance; re-computes the centroids of each cluster and repeats the previous steps until the moment each

individual belongs to a stable cluster and, therefore, it is not possible to carry out more individual transfers among cluster. The method presents the disadvantage of limiting the search for an optimal partition of individuals since it restricts the number of predefined clusters. Still, it has the advantage of defining the regions belonging to each cluster as well as the distance of each group.

4. Empirical Application

The empirical application of the cluster analysis present in this research work follows the methodological steps suggested in the cluster analysis literature.

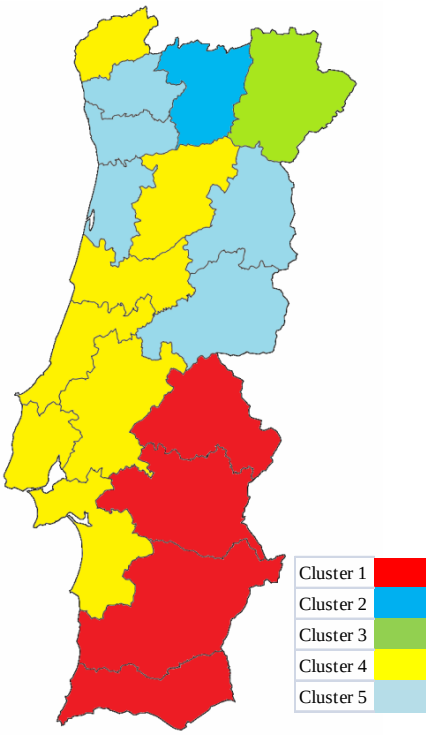
First of all the observations in analysis where selected. Considering the availability of the number of register unemployment individuals by municipalities, and its characterization throughout a range of distinct variables, was possible to aggregate the municipality observations by district - the object of study in this research work. Such geographical option is explained by administrative reasons concerning the political and social cohesion of these geographical areas.

In a second step was made the selection of variables that offer the required information for the districts clustering. The selected variables comprise individual and labour characteristics. Regarding the individual characteristics is available information related to gender, age and education of the register unemployed individuals. Regarding labour characteristics is known the duration of the unemployment register period and the unemployed individual experience concerning the labour market (looking for a first employment *versus* looking for a new employment). Since the variables are defined in different measurement scales it was important to transform them in such a way that they could appear with the same standard measurement scale. This transformation avoids that some of the measures of similarity/distance reflect the weight of the variables with bigger absolute values and dispersion. In particular the standardization applied measures the relative weight of each variable in the total of the unemployed individuals register in the public employment offices, by district. This is, was computed the ratio of registered unemployed, by variable, regarding the total number of registered unemployed individuals.

This k-means algorithm assumes that the number of groups (clusters) is previously known. However such assumption is rather unrealistic for most analysis problems since, normally, the data properties are not known. Thus the estimation of the number k of clusters is a frequent problem not only when applying the k-means algorithm but also other methods of clusters generation.In Stata, two testing methods for the optimum number of clusters were consider. Both Calinski-Harabasz pseudo F-test and Duda-Hart pseudo test point out for 5, 7 or more clusters. Following a parsimony rule, to avoid excessive information lost, the option was to divide the districts in 5 clusters.

The results for the 5-means methodology are presented in Figure 1.

Figure 1. Spatial distribution of districts by cluster



The analysis of the Portuguese territory allows concluding that exist a clear distinction between the north and south districts and between littoral (more urban) and inland (more rural) districts.

The visual results do not clarify, however, which is the mean profile of a registered unemployed individual. This mean profile was obtained through the estimation, in each cluster, of the mean values for the variables. The aim is to understand which variables better characterized the unemployed individual register in the public employment services of

each district.

The mean profiles of the registered unemployed individuals, for each cluster, are described in Table 2.

Table 2. Average proportion of each variable, by cluster (%)

	Cluster	1	2	3	4	5
Gender	Male	48,6	47,0	43,2	47,4	44,7
	Female	51,4	53,0	56,8	52,6	55,3
Age	<25 years	16,4	15,7	15,2	13,1	13,4
	25-34 years	26,3	23,5	24,5	25,3	21,6
	35-54 years	41,8	43,9	41,9	43,7	45,2
	>=55 years	15,6	16,9	18,4	17,9	19,8
Education	Less than 1º CB	8,1	6,7	9,0	5,0	5,6
	1º CB	24,4	32,5	27,4	25,2	32,9
	2º CB	18,7	19,1	20,4	18,8	18,4
	3º CB	20,5	16,8	17,2	21,2	19,1
	Secondary	21,0	16,3	15,4	20,1	15,5
	Superior	7,3	8,7	10,6	9,6	8,5
Position in the labour market	First Employment	9,1	16,9	17,6	7,9	9,4
	New Employment	90,9	83,1	82,4	92,1	90,6
Unemployment duration	Less than 1 year	76,0	62,9	69,3	69,6	61,4
	More than 1 year	24,0	37,1	30,7	30,4	38,6

5. Cluster unemployment profiles

Cluster 1: *Beja, Évora, Faro* and *Portalegre*. This cluster is characterized by the fact that it is the one with the biggest proportion of unemployed men and young individuals with an age inferior to 34 years. Maybe due to the youth of its registered unemployed individuals it is the cluster with higher levels of formal education and lower duration spells. The national increase of the number of registered unemployed individuals should have had as consequence the increase of proportion of individuals looking for a new employment after losing their ties with the formal labour market.

Cluster 2: *Vila Real*. The cluster presents specificities that transform the region in a particular one. It is the district with the bigger proportion of unemployed individuals with age between 25 and 34 years and the bigger proportion of unemployed individuals with the first basic schooling cycle. It is also one of the regions with the bigger proportion of individuals looking for a first employment but with unemployment spells bigger than 1 year.

Cluster 3: *Bragança*. Like the district of *Vila Real, Bragança* appears in 2009 with special features. The rise in the broad registered unemployment stressed the proportional number of unemployed women with age superior to 35 years. It is the region in which the proportion of those without any formal education is higher but it is also the region where it is possible to find the bigger proportion of registered individuals with an university formal education. In 2009 it is also possible to notice that *Bragança* becomes the district with the bigger proportion of individuals looking for a first employment most of them with unemployment duration spell lower than 1 year.

Cluster 4: *Coimbra, Leiria, Lisboa, Santarém, Setúbal, Viana do Castelo* and Viseu. This group includes the districts with the lower proportion of unemployed young individuals (those aged less than 25 years). With a close relation to the previous fact, it should be noticed that it is the cluster with the bigger proportion of unemployed individuals looking for a new employment. The cluster aggregates regions close to the littoral and in the centre of the country suggesting that the changes in the number of registered unemployed had similar consequences in the unemployment profile of this contiguous geographical region.

Cluster 5: *Aveiro, Braga, Castelo Branco, Guarda* and *Porto*. Even if the cluster geographical dispersion is large, the five districts share among themselves the following characteristics. In 2009 they lose, for the benefit of *Bragança*, the bigger proportion of unemployed women even if the difference between male and female unemployment is the biggest in the whole country. It is the cluster with the bigger proportion of registered individuals aged more than 55 years and also the one with lower levels of schooling among the unemployed population. Due to the previous characteristics it is not strange to

observe that this is the group of districts with the higher proportion of long duration unemployed individuals (individuals with unemployment spell longer than 1 year).

6. Final Remarks

From the results analysis the main conclusion to withdraw is that there is a clear distinction among different regions of the country. This is particularly true when we analyse the littoral and the inland regions and the north and the south of the country. It is possible to identify different unemployment regional profiles according to the geographical areas. Other studies, applying the same methodology here adopted, had reached the conclusion that the littoral regions present development features different from the ones that could be found in inland areas. Those studies have alerted that any regional development policy should pay attention to this reality [6]. The previous mentioned authors advise to the particular fact that any regional development policy should prevent the youth migration from the inland (rural areas) to the littoral (urban areas) throughout the creation of employment opportunities in the less development areas of the country. This research work stresses that conclusion and highlights this fact.

The second conclusion to withdraw is that there is a window of opportunities to implement different labour market public policies in the Portuguese territory. It was possible to find regions where the youth unemployment, the long-term unemployment or the female unemployment are concentrated. There are also regions characterised by a registered unemployed population with low levels of formal education, without any previous connection to the labour market or, in opposition, with previous links to labour market experiences.

7. References

[1] M. Arandarenko and M. Juvicic, „Regional Labour Market Differences in Serbia: Assessment and Policy Recommendations", The European Journal of Comparative Economics, 4(2), 2007, pp. 299-317.

[2] A. G. Babucea, "Using Cluster Analysis for Studying the Proximity of Registered Unemployment at the Level of Counties at the Beginning of the Economic Crisis", Annals of the "Constantin Brâncusi", University of Târgu Jiu, Economic series, Issue 1/2009

[3] P. Giudici, and S. Figini, "Applied Data Mining for Business and Industry" (2nd Ed). London, John Wiley & Sons Ltd., 2009

[4] E. Marelli, "Specialization and Convergence of European Regions", The European Journal of Comparative Economics, 2, 2007, pp. 149-178.

[5] D. Nadiya, "Econometric and Cluster Analysis of Potential and Regional Features of the Labor Market of Poland", Ekonomia, 21, 2008, pp. 28-44.

[6] J. O. Soares, M. M. L. Marques and C. M. F. Monteiro, „A Multivariate Methodology to Uncover Regional Disparities: A Contribution to Improve European Union and Governmental Decisions", European Journal of Operational Research, 145, 2003, pp. 121-135.

[7] R. C. Tryon, "Cluster Analysis", New York, McGraw-Hill, 1939