Investigação Operacional 2013

Atas do

XVI Congresso

da Associação Portuguesa
de Investigação Operacional

Bragança
3 a 5 de junho de 2013

Editado por
José F. Oliveira
Clara B. Vaz

Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Bragança

Atas do XVI Congresso da Associação Portuguesa de Investigação Operacional

Editores:

José Fernando Oliveira

Clara Bento Vaz

Com a colaboração de:

Ana Isabel Pereira

Instituto Politécnico de Bragança
3 a 5 de junho 2013

Este volume contém artigos submetidos e apresentados no XVI Congresso da Associação Portuguesa de Investigação Operacional, realizado em Bragança, Portugal, de 3 a 5 de junho de 2103.

Título: **Livro de Atas do XVI Congresso da Associação Portuguesa de Investigação Operacional**

Editores:
José Fernando Oliveira
Clara Bento Vaz

Colaboração:
Ana Isabel Pereira

Primeira edição, em formato eletrónico, junho 2013
ISBN: 978-972-745-154-8

Comissão de Programa

José Fernando Oliveira (Presidente), Universidade do Porto, Faculdade de Engenharia
Agostinho Agra, Universidade de Aveiro, Departamento de Matemática
Ana Isabel Pereira, Instituto Politécnico de Bragança, Escola Superior de Tecnologia e Gestão
Ana Paula Teixeira, Universidade de Trás-os-Montes e Alto Douro, Departamento de Matemática e CIO
Ana Viana, Instituto Politécnico do Porto, Instituto Superior de Engenharia
Clara Bento Vaz, Instituto Politécnico de Bragança, Escola Superior de Tecnologia e Gestão
Filipe Alvelos, Universidade do Minho, Escola de Engenharia
Isabel Gomes, Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia
João Luís Soares, Universidade de Coimbra, Faculdade de Ciências e Tecnologia
Joaquim Borges Gouveia, Universidade de Aveiro, Dep. de Economia, Gestão e Engenharia Industrial
Jorge Orestes Cerdeira, Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia
José Manuel Valério de Carvalho, Universidade do Minho, Escola de Engenharia
Margarida Vaz Pato, Universidade Técnica de Lisboa, Instituto Superior de Economia e Gestão
Maria Antónia Carravilla, Universidade do Porto, Faculdade de Engenharia
Maria Eugénia Captivo, Universidade de Lisboa, Faculdade de Ciências e CIO
Maria João Alves, Universidade de Coimbra, Faculdade de Economia
Marília Pires, Universidade do Algarve, Faculdade de Ciências e Tecnologia
Miguel Constantino, Universidade de Lisboa, Faculdade de Ciências
Mónica Oliveira, Universidade Técnica de Lisboa, Instituto Superior Técnico
Susana Relvas, Universidade Técnica de Lisboa, Instituto Superior Técnico

Comissão Organizadora

Clara Bento Vaz, Instituto Politécnico de Bragança (Presidente)
Ana Isabel Pereira, Instituto Politécnico de Bragança
António Jorge Trindade Duarte, Instituto Politécnico de Bragança
Carla Alexandra Soares Geraldes, Instituto Politécnico de Bragança
Carla Maria Carneiro Alves, Instituto Politécnico de Bragança
Carla Sofia Renca da Cruz, Instituto Politécnico de Bragança
Carla Sofia Veiga Fernandes, Instituto Politécnico de Bragança
Carlos Jorge da Rocha Balsa, Instituto Politécnico de Bragança
Edite Martins Cordeiro, Instituto Politécnico de Bragança
Elisa Margarida Correia de Barros, Instituto Politécnico de Bragança
Florbela Alexandra Pires Fernandes, Instituto Politécnico de Bragança
Francisco José Pires Peito, Instituto Politécnico de Bragança
Ilda Marisa de Sá Reis, Instituto Politécnico de Bragança
João Paulo Pais de Almeida, Instituto Politécnico de Bragança
José Fernando Oliveira, Universidade do Porto, Faculdade de Engenharia
José Mário Escudeiro de Aguiar, Instituto Politécnico de Bragança
José Paulo Macedo Matias, Instituto Politécnico de Bragança
Maria de Fátima Silva Pacheco, Instituto Politécnico de Bragança
Maria Prudência Gonçalves Martins, Instituto Politécnico de Bragança
Paula Maria Pereira de Barros, Instituto Politécnico de Bragança

Índice

A Comparative Study of Two Optimization Clustering Techniques on Unemployment Data

Elisa Barros*, Alcina Nunes*, Carlos Balsa*

* *Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal*
E-mail: ebarros@ipb.pt , alcina@ipb.pt , balsa@ipb.pt

Abstract

An important strategy for data classification consists in organising data points in clusters. The k-means is a traditional optimisation method applied to cluster data points. Using a labour market database, we suggest the application of an alternative method based on the computation of the dominant eigenvalue of a matrix related with the distance among data points. This approach presents results consistent with the results obtained by the k-means.

Key words: Clustering methods, k-means, spectral clustering, unemployment data mining

1 Introduction

Clustering is an important process for data classification that consists in organising a set of data points into groups, called clusters. A cluster is a subset of an original set of data points that are close together in some distance measure. In other words, given a data matrix containing multivariate measurements on a large number of individuals (observations or points), the aim of the cluster analysis is to build up some natural groups (clusters) with homogeneous properties out of heterogeneous large samples [Kaufman and Rousseeuw, 1990].

Groups are based on similarities. The similarity depends on the distance between data points and a reduced distance indicates that they are more similar. Several distinct methods can be used to measure the distance among the elements of a data set. Along this work we will consider the traditional Euclidian distance, i.e., the 2-norm of the differences between data points vectors.

There are two main classes of clustering techniques: hierarchical and optimization methods. In hierarchical clustering is not necessary to know in advance the number of subsets in which we want to divide the data. The observations are successively included in groups of different dimensions depending on the level of clustering. The result is a set of nested partitions. In each step of the process, two groups are either merged (agglomerative methods) or divided (divisive methods) according to some criteria [Matinez et al, 2011]. In the agglomerative approach, single-members clusters (clusters with only one observation) are increasingly fused until all observations are in only one cluster. The divisive approach starts with a single set containing all points. This group will be increasingly divided as the distance between points is reduced. The set of nested partitions is represented graphically by a dendrogram that has a tree shape indicating the distance′s hierarchical dependence. The dendrogram can help to identify the number of clusters that should be considered during the partition of the data set. After that a partition method like k-means can be applied to identify each cluster.

The k-means [MacQueen, 1967] is an optimization method that partitions the data in exactly k clusters, previously determine. This is achieved in a sequence of steps which begins, for instance, with an initial partition randomly generated. In each step the clusters centroid (arithmetic vector mean) is computed. The minimum distance between each data point and the clusters′ different centroids will decide the formation of new clusters. The formation of a new cluster implies assigning each observation to the cluster which presents the lowest distance. After that the centroids are (re)calculated and the former step is repeated until the moment each individual belongs to a stable cluster, i.e., when the sum of the squared distances to the centroid of all data point over all the clusters is minimized. The algorithm presents a rather fast convergence, but one cannot guarantee that the algorithm finds the global minimum [EldÃ©n, 2007].

Spectral clustering is also an optimization method. This method is becoming very popular in recent years because it has been included in algorithms used in the identification of the human genome or in web

browsers. Beyond biology and information retrieval the method has other fields of application such as image analysis and, in some cases, it can perform better than standard algorithms such as k-means and hierarchical clustering [Matinez et al, 2011]. Spectral clustering methods use the k dominant eigenvectors of a matrix, called affinity matrix, based on the distance between the observations. The idea is grouping data points in a lower-dimensional space described by those k eigenvectors [Mouysset et al, 2008]. The approach may not make a lot of sense, at first, since we could apply the k-means methodology directly without going through all the matrix calculations and manipulations. However, some analyses show that mapping the points to this k-dimensional space can produce tight clusters that can easily be found applying k-means [Matinez et al, 2011].

In the present research work, spectral clustering is applied in an unusual context concerning the traditional data mining analysis. We classify 278 Portuguese mainland municipalities (*concelhos*) regarding the type/characteristics of unemployment official registers. The set of observations, $x_1, \ldots, x_{278}$, that contains 278 vectors, whose 11 coordinates are the values for some of the indicators used to characterise Portuguese unemployment (gender, age classes, levels of formal education, situation relating unemployment and unemployment duration), is divided in k clusters. The classification of observations resulting from the spectral method is than compared to the classification given by the traditional k-means method. The results are analysed from both mathematical and economic points of view. The main goal is to find evidence regarding which method produces the best cluster partition and, accordingly, to understand if the resulting clusterisation makes sense in terms of the spatial distribution of unemployment characteristics, over a country administrative territory. The idea is to understand if a particular cluster methodology for data mining analysis provides useful and suitable information that could be used to the development of national, regional or local unemployment policies.

The paper is divided as follows. The k-means method and the spectral clustering method are presented in sections 2 and 3, respectively. The methods description is followed by section 4 where data and variables analysed are also presented and described. In section 5 we move ahead toward the optimal number of clusters applying both selected methods. In section 6 the results are presented and discussed. Our concluding remarks can be found on section 7.

2 The k-means method

We are concerned with m data points $x_i \in \mathbb{R}^n$ that we want classify in k clusters, where k is predetermined. We organize the data as lines in a matrix $X \in \mathbb{R}^{m \times n}$. To describe the k-means method as proposed in [EldÃ©n, 2007] we denote a partition of vectors $x_1, \ldots, x_m$ in k clusters as $\prod = \{\pi_1, \ldots, \pi_k\}$ where

$$\pi_j = \{\ell : x_\ell \in \text{cluster} j\}$$

defines the set of vectors in cluster j. The centroid, or the arithmetic mean, of the cluster j is:

$$m_j = \frac{1}{n_j} \sum_{\ell \in \pi_j} x_\ell \tag{1}$$

where n_j is the number of elements in cluster j. The sum of the squared distance, in 2-norm, between the data points and the j cluster′s centroid is known as the *coherence*:

$$q_j = \sum_{\ell \in \pi_j} \|x_\ell - m_j\|_2^2 \tag{2}$$

The closer the vectors are to the centroid, the smaller the value of q_j. The quality of a clustering process can be measured as the *overall coherence*:

$$Q\left(\prod\right) = \sum_{j=1}^{k} q_j \tag{3}$$

The k-means is considered an optimization method because it seeks a partition process that minimizes $Q(\prod)$ and, consequently, finds an optimal coherence. The problem of minimizing the *overall coherence* is NP-hard and, therefore, very difficult to achieve. The basic algorithm for k-means clustering is a two step heuristic procedure. Firstly, each vector is assigned to its closest group. After that, new centroids are computed using the assigned vectors. In the following version of k-means algorithm, proposed by [EldÃ©n, 2007], these steps are alternated until the changes in the *overall coherence* are lower than a certain tolerance previously defined.

The k-means algorithm

1. Start with an initial partitioning $\prod^{(0)}$ and compute the corresponding centroid vectors $m_j^{(0)}$ for $j = 1, \ldots, k$. Compute $Q(\prod^{(0)})$. Put $t = 1$.
2. For each vector x_i find the closest centroid. If the closest centroid is m_p^{t-1} assign x_i to $\pi_p^{(t)}$.
3. Compute the centroids $m_j^{(t)}$ for $j = 1, \ldots, k$ of the new partitioning $\prod^{(t)}$.
4. If $\left|Q(\prod^{(t)}) - Q(\prod^{(t-1)})\right| < \text{tol}$, stop; Otherwise $t = t + 1$ and return to step 2.

Since it is an heuristic algorithm there is no guarantee that k-means will converge to the global minimum, and the result may depend on the initial partition $\prod^{(0)}$. To avoid this issue, it is common to run it multiple times, with different starting conditions choosing the solution with the smaller $Q(\prod)$.

3 Spectral clustering method

Let $x_1, \ldots, x_m$ be a m data points set in a n-dimensional euclidian space. We want to group these m points in k clusters in order to have better within-cluster affinities and weaker affinities across clusters. The affinity between two observations x_i and x_j is defined by [Ng et al, 2002] as:

$$A_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right) \tag{4}$$

where σ is a scaling parameter that determines how fast the affinity decreases with the distance between x_i and x_j. The appropriate choice of this parameter is crucial [Matinez et al, 2011]. In [Ng et al, 2002] we can find a description of a method able to choose the scaling parameter automatically.

The spectral clustering algorithm proposed by [Ng et al, 2002] is based on the extraction of dominant eigenvalues and their corresponding eigenvectors from the normalized affinity matrix $A \in \mathbb{R}^{m \times m}$. The components A_{ij} of A are given by equation 4, if $i \neq j$, and by $A_{ii} = 0$, if $i = j$. The sequence of steps in the spectral clustering algorithm is presented as follows:

The spectral clustering algorithm

1. Form the affinity matrix A as indicated in equation 4.
2. Construct the normalized matrix $L = D^{-1/2} A D^{-1/2}$ with $D_{ii} = \sum_{j=1}^{m} A_{ii}$.
3. Construct the matrix $V = [v_1 v_2 \ldots v_k] \in \mathbb{R}^{m \times k}$ by stacking the eigenvectors associated with the k largest eigenvalues of L.
4. Form the matrix Y by normalizing each row in the $m \times k$ matrix V (i.e. $Y_{ij} = V_{ij} / \left(\sum_{j=1}^{k} V_{ij}^2\right)^{1/2}$).
5. Treat each row of Y as a point in $\mathbb{R}^k$ and group them in k clusters by using the k-means method.
6. Assign the original point x_i to cluster j if and only if row i of matrix Y was assigned to cluster j.

4 Data description

The 278 data points represents the Portuguese continental *concelhos*. Each data point have 11 coordinates representing characteristics of the unemployed register individuals. Indeed, the unemployed individuals registered in the Portuguese public employment services of the *Instituto de Emprego e FormaÃ§Ã£o Profissional (IEFP)* present a given set of distinctive characteristics related with gender, age, formal education, unemployment spell (unemployment for less than a year or more than a year)and situation related with the unemployment situation (unemployed individual looking for a first employment or for another employment). These characteristics are important determinants of unemployment and are important economic vectors regarding the development of public employment policies. Well targeted policies are more efficient, in terms of expected results, and avoid the waste of scarce resources.

A complete study of regional similarities (or dissimilarities) in a particular labour market, as the Portuguese, should not be limited by a descriptive analysis of the associated economic phenomena. It should also try to establish spacial comparison patterns among geographic areas in order to develop both national and regional public policies to fight the problem. Indeed high unemployment indicators and regional inequalities are major concerns for European policy-makers since the creation of European Union. However, even if the problem is known the policies dealing with unemployment and regional inequalities have been few and weak [Overman and Puga, 2002]. In Portugal, in particular, there are

some studies that try to define geographic, economic and social homogeneous groups [Soares et al, 2003]. To the best of our knowledge, there are no studies that offer an analysis of regional unemployment profiles. Other economies are starting to develop this kind of statistical analysis using as a policy tool the cluster analysis methodology [Arandarenko and Juvicic, 2007, López-Bazo et al, 2005, Nadiya, 2008].

The data concerning the above mentioned characteristics are openly available in a monthly period base in the website of *IEFP* (http://www.iefp.pt/estatisticas/Paginas/Home.aspx). Additionally, the month of December gives information about the stock of registered unemployed individuals at the end of the respective year. In the case of this research work, data from unemployment registers in 2012 have been used. The eleven variables available to characterise the individuals and that have been used here are divided in demographic variables and variables related with the labour market. These variables are dummy variables, measured in percentage of the total number of register individuals in a given *concelho*, and describe the register unemployed as follows: 1: Female, 2: Long duration unemployed (individual unemployed for more than 1 year), 3: Unemployed looking for a new employment, 4: Age lower than 25 years, 5: Age between 25 and 35 years, 6: Age between 35 and 54 years, 7: Age equal or higher than 55 years, 8: Less than 4 years of formal education (includes individuals with no formal education at all), 9: Between 4 and 6 years of formal education, 10: Between 6 and 12 years of formal education and 11: Higher education.

Women, individuals in a situation of long duration unemployment, younger or older unemployed individuals and the ones with lower formal education are the most fragile groups in the labour market and, consequently, are the most exposed to unemployment situations. They are also the most challenging groups regarding the development of public employment policies.

5 Toward the optimal number of clusters

We begin by applying the k-means method to partition in k clusters the data points set $x_1, \ldots, x_m$, with $m = 278$ Portuguese mainland *concelhos* regarding the 11 chosen unemployment characteristics. As the optimal number of targeted groups is unknown *a priori*, we repeat the partition for $k = 2,\ 3,\ 4$ and 5 clusters.

To evaluate the quality of the results from the cluster methodology and to estimate the correct number of groups in our data set we resort the silhouette statistic framework. The silhouette statistic introduced by [Kaufman and Rousseeuw, 1990] is a way to estimate the number of groups in a data set. Given observation x_i, the average dissimilarity to all other points in its own cluster is denoted as a_i. For any other cluster c, the average dissimilarity of x_i to all data points in cluster c is represented by $\bar{d}(x_i, c)$. Finally, b_i denote the minimum of these average dissimilarities $\bar{d}(x_i, c)$. The *silhouette width* for the observation x_i is:

$$s_i = \frac{(b_i - a_i)}{\max\{b_i,\ a_i\}}. \tag{5}$$

The *average silhouette width* is obtained by averaging the s_i over all observations:

$$\bar{s}_i = \frac{1}{m}\sum_{i=1}^{m} s_i. \tag{6}$$

If the *silhouette width* of an observation is large it tends to be well clustered. Observations with small *silhouette width* values tend to be those that are scattered between clusters. The *silhouette width* s_i in Equation 5 ranges from -1 to 1. If an observation has a value close to 1, then it is closer to its own cluster than it is to a neighbouring one. If it has a *silhouette width* close to -1, then it is a sign that it is not very well clustered. A *silhouette width* close to zero indicates that the observation could just as well belong to its current cluster or one that is near to it.

The *average silhouette width* (equation 6) can be used to estimate the number of clusters in the data set by using the partition with two or more clusters that yield the largest average silhouette width [Kaufman and Rousseeuw, 1990]. As a rule of thumb, it is considered that an *average silhouette width* greater than 0.5 indicates a reasonable partition of the data, and a value less than 0.2 would indicate that the data do not exhibit a cluster structure [Matinez et al, 2011].

Figure 1 presents the *silhouette width* corresponding to the case of four different partitions of the data points set, this is, $k = 2,\ 3,\ 4$ and 5 clusters resulting from the application of the k-means method.

As it is possible to observe, the worst cases occur, clearly, when $k = 3$ and $k = 5$. For these cases, some clusters present negative values and others appear with small (even if positive) silhouette indexes.

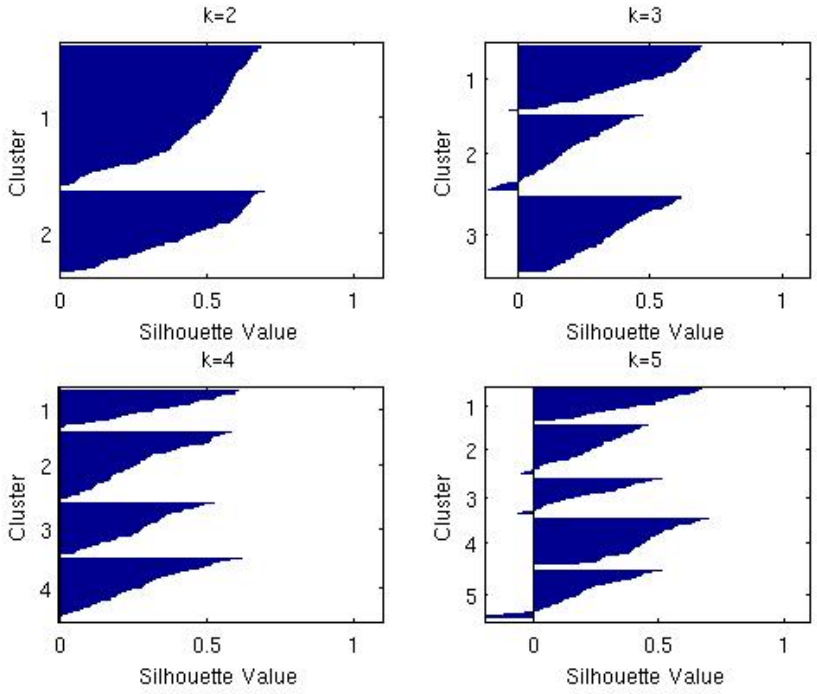

Figure 1: *Silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the k-means method.

In the case of $k = 2$ and $k = 4$ clusters there are no negative values, however we find large silhouette values mostly in the case of the two clusters partition.

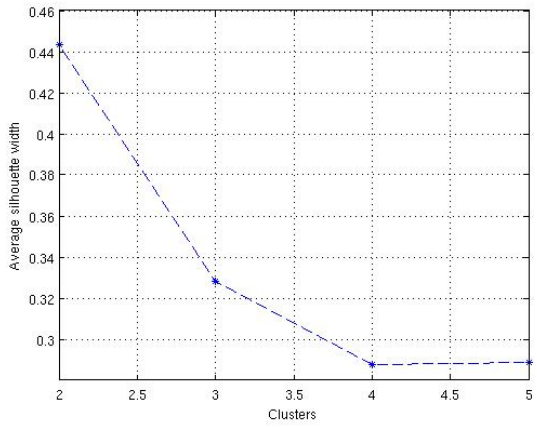

Figure 2: *Average silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the k-means method.

To get a single number that is able to summary and describe each clustering process, we find the *average of the silhouette* values (equation 6) corresponding to $k = 2,\ 3,\ 4$ and 5. The results can be observed in figure 2.

The two cluster solution presents an average silhouette value near 0.44 and the four cluster solution presents an average silhouette value near 0.29. These results confirm the ones above. The best partition obtained with the application of the k-means method occurs with $k = 2$. Nonetheless, the *average of the silhouette* is close but smaller than 0.5 which reveals that the data set does not seem to present a strong trend to be partitioned in two clusters.

Figure 3 shows the *silhouette width* corresponding to each observation in the case of four different partitions of the data set points. This is, in $k = 2,\ 3,\ 4$ and 5 clusters, resulting from the application of the spectral clustering method.

In this case all the tested partitions present clusters where can be observed negative values. The worst

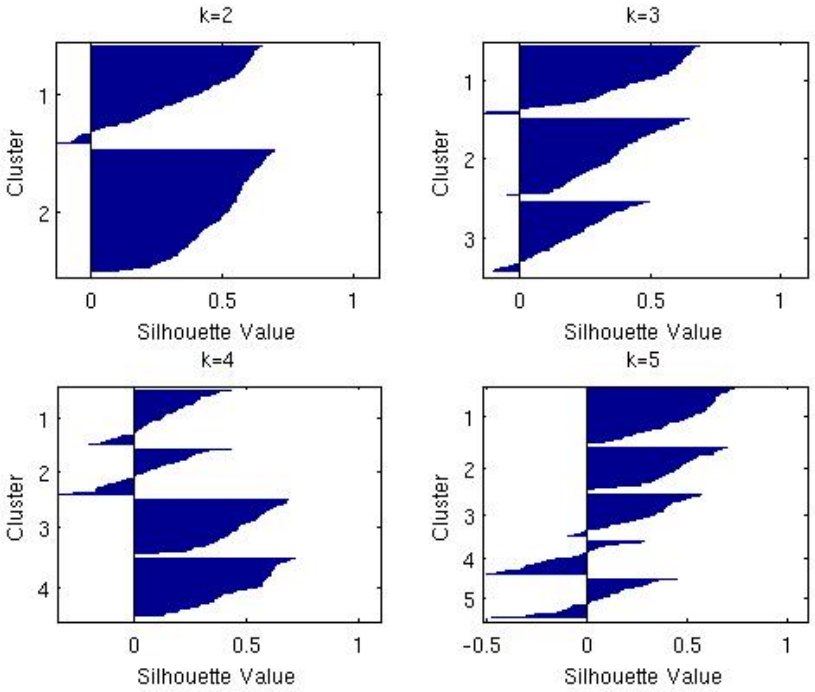

Figure 3: *Silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the spectral method.

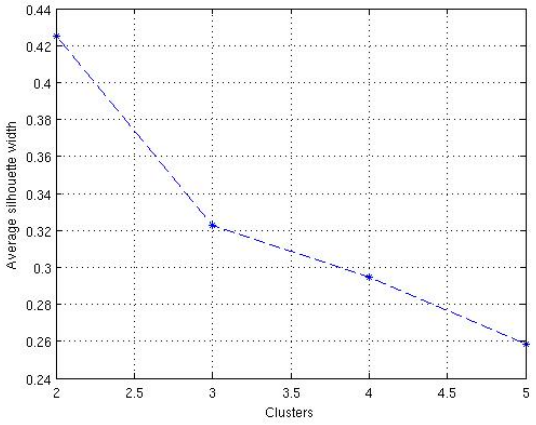

Figure 4: *Average silhouette width* for $k = 2,\ 3,\ 4$ and 5 clusters resulting from the spectral method.

cases occur, clearly, when $k = 4$ and $k = 5$. Here we get values close to -0.5. In the case $k = 3$ is possible to observe negative values in the three cluster obtained whereas in the case of $k = 2$ the negative values are just observed in one of the two clusters.

The trend observed with the *silhouette width* is confirmed by the *average of the silhouette* values corresponding to the spectral clustering process with $k = 2,\ 3,\ 4$ and 5 clusters (figure 4).

The two cluster solution has an average silhouette value near 0.43 and decrease as the number of clusters increases. The best partition obtained with the spectral clustering method occurs with $k = 2$. These results are in agreement with the partitioning found by using the k-means method. The *average of the silhouette* value (0.43) is very close to the one calculated with k-means method (0.44).

As mentioned before, the results obtained with the k-means method agree with the results obtained with the application of the spectral methods. The best partition of the data set is accomplished with two clusters. However, this trend is not completely crystal clear. Indeed, the *average of the silhouette* in the two cases is smaller than 0.5. The computed value indicates that the distance between the two considered clusters is not very large. This conclusion can be visually confirmed by the hierarchical distance between two data points. This distance is illustrated by the dendrogram presented on figure 5. In the dendrogram

we can observe an initial bifurcation that divides the data in two main groups, but the distance between them is near 0.6. We can see also that one of the two main clusters can be partitioned into other two if we consider a distance near 0.5, very close to the distance between the two main clusters.

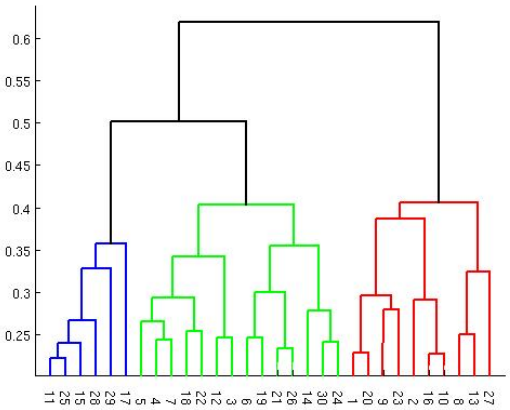

Figure 5: Dendrogram.

6 Mathematical and economic results′ analysis

Both spectral clustering method and k-means method indicate that the data are best partitioned into two clusters. The statistical properties of theses two clusters are presented in table 1.

method	j	n_j	q_j	Q
k-means	1	177	3.4161	5.3276
	2	101	1.9115	
spectral	1	154	2.7511	5.3536
	2	124	2.6026	

Table 1: Statistical properties of the two clusters resulting from k-means and spectral methods

Despite the number of observations in each cluster is not the same, it appears that for both methods the first cluster is the largest. This is, includes a bigger number of *concelhos*: $n_1 = 177$ for the k-means and $n_1 = 154$ for the spectral method. The difference of 23 observations for the first cluster is reflected in the computed local coherence q (equation 2) that is larger for the k-means methods ($q_1 = 3.4161$). The second cluster comprises $n_2 = 101$ observations and presents a local coherence of $q_2 = 1.9115$, for the k-means, and $n_2 = 124$ observations and a local coherence of $q_2 = 2.6026$ for the spectral method. Although the differences between the computed coherence for each cluster, we can observe that both methods achieve a very similar overall coherence (equation 3), $Q \approx 5.3$ for the k-means and $Q \approx 5.4$ for the spectral method.

For a more complete comparison analysis of the results obtained by k-means and spectral methods, it is also important to analyse two distribution measures: mean and standard deviation. The measures are presented for each one of the 11 variables used in the cluster analysis. In figure 6 we compare the mean value obtained for the 11 parameters that characterise the two clusters obtained by the two clusterisation methods. In figure 7 we compare the standard deviation value. Note that in these two figures the comparison analysis is done regarding the cluster methods applied.

It is visible that the computed mean values, regarding each one of the variables, are very similar in the two clusters independently of the cluster method used. For the computed standard deviation values we can observe a first cluster where the standard deviation, for the overall set of variables, are slightly higher for the k-means and a second cluster where the observed trend is reversed. In short, we can observe that the results for both methods are similar regarding the measure of central tendency of each one of the variables but the variability of values, regarding the central tendency, differ between cluster methods.

The mean and standard deviation measures can be compared regarding the values computed by cluster. From this point of view the analysis would have an economic focus. So, in figure 8 we compare

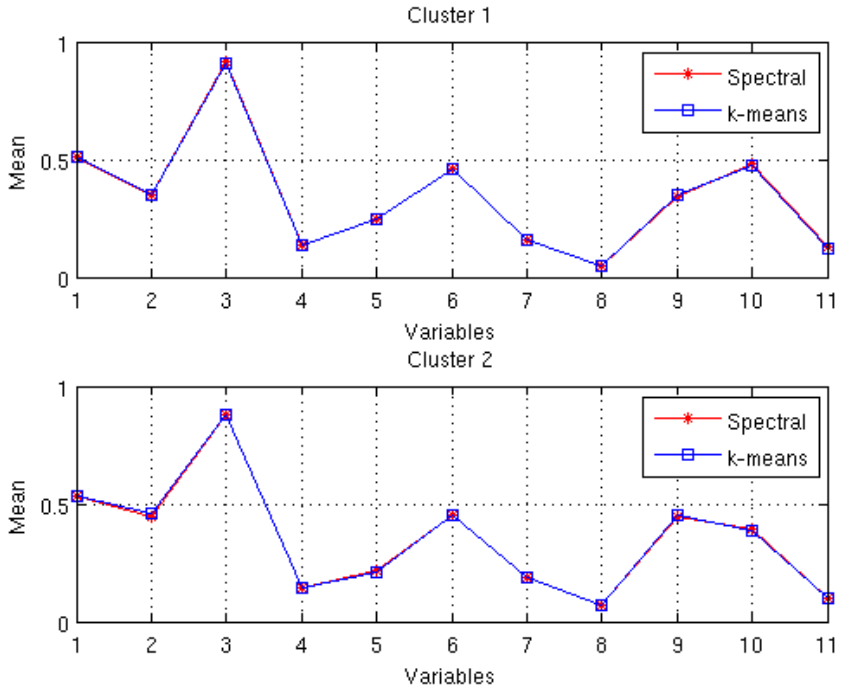

Figure 6: Mean values computed for the two clusters methods by cluster.

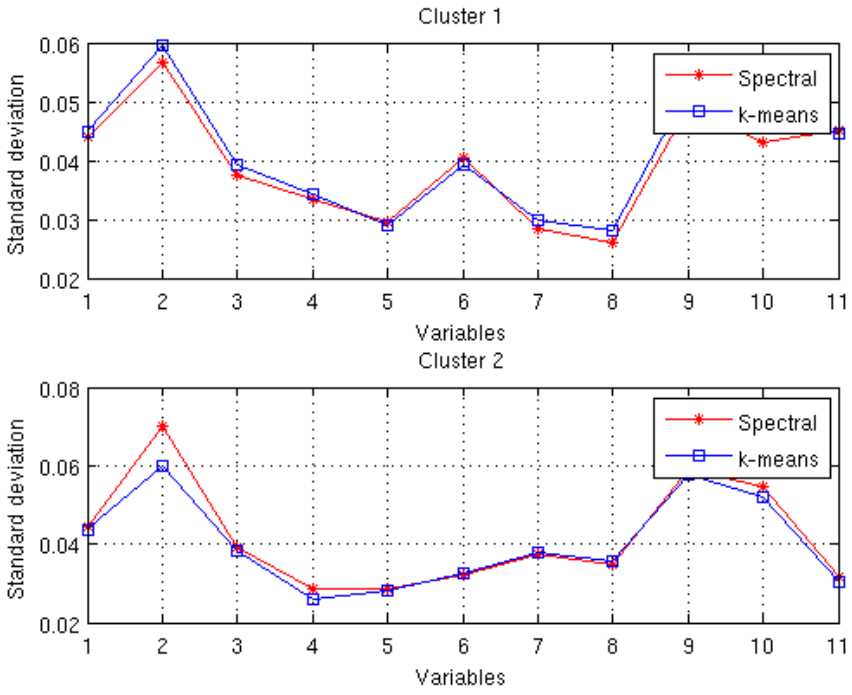

Figure 7: Standard deviation values computed for the two clusters methods by cluster.

the mean value obtained for the 11 parameters for each one of the clusters by cluster method. In figure 9 we compare the computed standard deviation value.

From the figure 8 and figure 9 it is possible to observe that both methods retrieve clusters that present the same pattern. In the second cluster (cluster 2) are gathered the Portuguese mainland *concelhos* that present a higher percentage of unemployed register individuals with more problematic characteristics - women, long duration unemployed individuals, individuals that are looking for a job for the first time

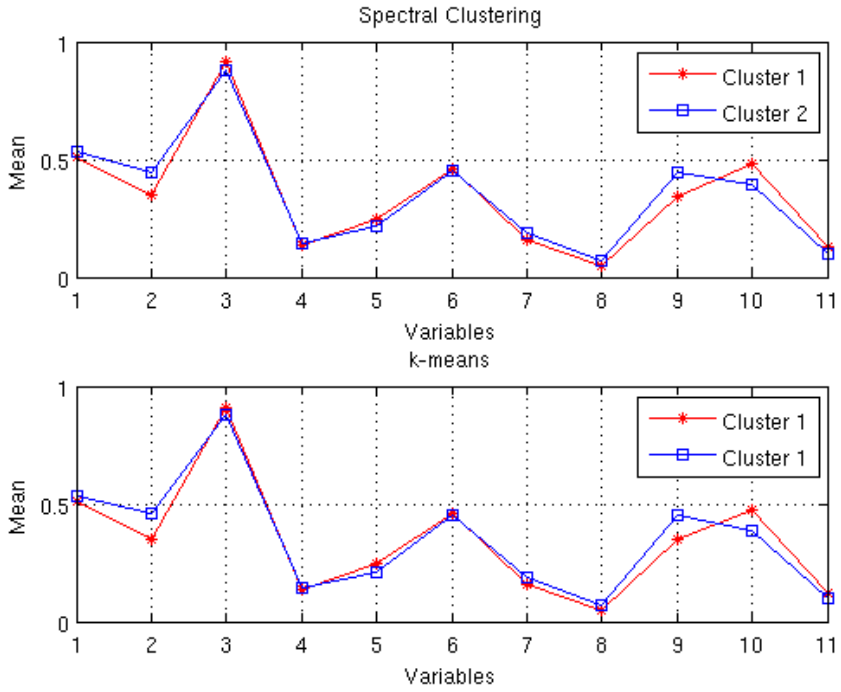

Figure 8: Mean values computed for the two clusters resulting from k-means and spectral method.

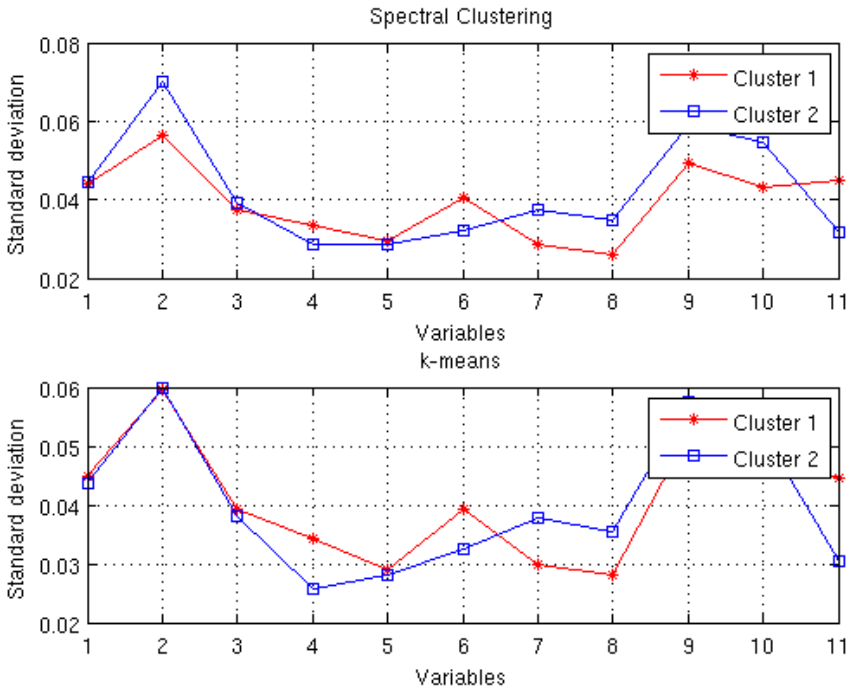

Figure 9: Standard deviation values computed for the two clusters resulting from k-means and spectral method.

(individuals with no connections with the labour market), individuals with more than 55 years and with lower number of years of formal education (for example, this cluster gathers the *concelhos* with a lower percentage of unemployed individuals with a higher education). As mentioned before these groups of individual are the most fragile labour market groups. Both cluster methods seem to divide the total number of *concelhos* in two economic meaningful clusters.

Regarding the standard deviation we observe that the k-means method retrieve clusters that present a lower variability among the observations in each cluster, by variable. The variability seems to be lower for the overall set of characteristics even if the k-means method divides the total number of observations in more uneven clusters.

7 Concluding remarks

In short, both methods denote the same data partition. Applying both methods, the data partition into two clusters minimises the dispersion of data values. The use of the spectral clustering method in an unusual economic application shows potential benefits. Without algorithm parameters refinement the method presented results that are consistent with the k-means results. From the economic point of view both methods show the importance of dividing Portuguese *concelhos* in well defined groups which could be object of distinct public policies. Well targeted labour market measures are, recognisable, more efficient with the cluster methodology helping the identification of different and well defined target regions.

References

[MacQueen, 1967] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1:281-297.

[Matinez et al, 2011] Martinez, W. L. and Martinez, A. R. and Solka J. L. (2011). Exploratory Data Analysis with MATLAB. *CRC Press.*

[EldÃ©n, 2007] EldÃ©n, L. (2007). Matrix Methods in Data Mining and Pattern Recognition. *SIAM.*

[Mouysset et al, 2008] Mouysset, S. and Noailles, J. and Ruiz, D. (2008). Using a Global Parameter for Gaussian Affinity Matrices in Spectral Clustering. *High Performance Computing for Computational Science - VECPAR 2008*, 378 - 390.

[Ng et al, 2002] Ng, A. Y. and Jordan, M. I. and Weiss Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS)*, 14:849-856.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. *New York: John Wiley & Sons.*

[Overman and Puga, 2002] Overman, H.G. and Puga, D. (2002). Unemployment clusters across Europe's regions and countries. *Economic Policy*, 17(34):115-148.

[Soares et al, 2003] Soares, J. O., Marques, M. M. L. and Monteiro, C. M. F. (2003). A Multivariate Methodology to Uncover Regional Disparities: A Contribution to Improve European Union and Governmental Decisions. *European Journal of Operational Research*, 45:121-135.

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

[López-Bazo et al, 2005] López-Bazo, E., Del Barrio, T. and Artís, M. (2005). Geographical Distribution of Unemployment in Spain. *Regional Studies*, 39(3):305-318.

[Nadiya, 2008] Nadiya, D. (2008). Econometric and Cluster Analysis of Potential and Regional Features of the Labor Market of Poland. *Ekonomia*, 21:28-44.