DOCUMENTS DE TREBALL

DE LA FACULTAT DE CIÈNCIES ECONÒMIQUES I EMPRESARIALS

Col·lecció d'Economia

Spanish unemployment: Normative *versus* analytical regionalisation procedures[*]

Juan Carlos Duque, Raúl Ramos, Manuel Artís

Grup d'Anàlisi Quantitativa Regional (Universitat de Barcelona)

Address for correspondence
Grup d'Anàlisi Quantitativa Regional
Espai of Recerca en Economia
Facultat of Ciències Econòmiques i Empresarials, Universitat of Barcelona
Avda. Diagonal 690 - 08034 Barcelona, Espanya
Tel: 934037043 Fax: 934021821
E-mail: jduque@eco.ub.es, rramos@ub.edu, manuel.artis@ub.edu

Received: March 2004

[*] Authors wish to thank E. López-Bazo, E. Pons and J. Suriñach for their helpful comments and suggestions to previous versions of this paper. The usual disclaimer applies. Financial support is gratefully acknowledged from CICYT SEC 2002-00165 project. Juan Carlos Duque also thanks the support of the Generalitat de Catalunya through the grant 2001FI 00296.

Abstract:

In applied regional analysis, statistical information is usually published at different territorial levels with the aim of providing information of interest for different potential users. When using this information, there are two different choices: first, to use normative regions (towns, provinces, etc.), or, second, to design analytical regions directly related with the analysed phenomena.

In this paper, provincial time series of unemployment rates in Spain are used in order to compare the results obtained by applying two analytical regionalisation models (a two stages procedure based on cluster analysis and a procedure based on mathematical programming) with the normative regions available at two different scales: NUTS II and NUTS I.

The results have shown that more homogeneous regions were designed when applying both analytical regionalisation tools. Two other obtained interesting results are related with the fact that analytical regions were also more stable along time and with the effects of scale in the regionalisation process.

Keywords: Unemployment, normative region, analytical region, regionalisation.

JEL Codes: E24, R23, C61.

Resumen:

En el análisis regional aplicado, la información estadística normalmente se encuentra disponible a diferentes niveles de desagregación territorial (o escalamiento) con el objetivo de proveer información a los diferentes usuarios potenciales. Cuando este tipo de información es analizada generalmente se tienen dos alternativas: La primera, consiste en utilizar las divisiones territoriales "normativas" (las oficialmente establecidas como pueblos, provincias, etc.), o, como segunda opción, diseñar regiones "analíticas" directamente relacionadas con el fenómeno analizado.

En este trabajo, series temporales sobre las tasas de desempleo en las provincias españolas son utilizadas con el objetivo de comparar los resultados obtenidos tras la aplicación de dos metodologías de regionalización analítica (aplicación de análisis cluster convencional en dos etapas y programación matemática) con las divisiones normativas disponibles a diferentes niveles de escalamiento: NUTS II y NUTS I.

Los resultados muestran como las regiones más homogéneas fueron diseñadas aplicando metodología analíticas. También destaca el hecho de que dichas regiones son más estables, en términos de homogeneidad, a lo largo del periodo analizado y para los diferentes escalamientos definidos.

Spanish unemployment: Normative *versus* analytical regionalisation procedures

1. INTRODUCTION AND OBJECTIVES

In applied regional analysis, statistical information is usually published at different territorial levels with the aim of providing information of interest for different potential users. When using this information, there are two different choices: first, to use *normative regions* (towns, provinces, etc.), or, second, to design *analytical regions* directly related with the analysed phenomena. This second option consists in the aggregation of territorial units of small size[1] without arriving at the upper level or, alternatively, in combining information from different levels[2].

In most cases, the aggregation of territorial information is usually done using "*ad-hoc*" criteria due to the lack of regionalisation methods with enough flexibility. In fact, most of these methods have been developed to deal with very particular regionalisation problems, so when they are applied in other contexts the results could be very restrictive or inappropriate for the considered problem. However, and with independence of the applied territorial aggregation method, there is an implicit risk, known in the literature as "Modifiable Areal Unit

[1] Apart from aspects such as the statistical secret or other legislation about the treatment of statistical data, according to Wise *et al*, (1997), this kind of territorial units are designed in such a way as to be above minimum population or household thresholds, to reduce the effect of outliers when aggregating data or to reduce possible inexactities in the data, and to simplify information requirements for calculations or to facilitate its visualisation and interpretations in maps.

[2] See, for example, Albert *et al*, (2003), who analyse the spatial distribution of economic activity using information with different levels of regional aggregation, NUTS III for Spain and France and NUTS II for the rest of countries, with the objective "using similar territorial units". López-Bazo *et al*. (1999) analyse inequalities and regional convergence at the European level in terms of GDP per capita using a database for 143 regions using NUTS II data for Belgium, Denmark, Germany, Greece, Spain, France, Italy, Netherlands and Portugal, and NUTS-I for the United Kingdom, Ireland and Luxemburg with the objective of ensuring the comparability of geographical units.

Problem" (Openshaw, 1984), and which is related with the sensitivity of the results to the aggregation of geographical data and its consequences on the analysis.

In this paper, provincial time series of unemployment rates in Spain are used in order to compare the results obtained by applying two analytical regionalisation models, each one representing a different regionalisation strategy: a two stages procedure based on cluster analysis and a procedure based on mathematical programming. The results will also be compared with normative regions available at two different scales: NUTS II and NUTS I.

The rest of the paper is organised in the following sections: Section 2 briefly describes the main characteristics of normative and analytical regions. Also the analytical regionalisation models used in the paper are presented. In section 3 the results of applying the two models in the context of provincial unemployment rates are shown with the aim of comparing normative and analytical regions, Last, most relevant conclusions are presented in section 4.

2. Normative vs. analytical regions: Regionalisation procedures

When analysing phenomena where the geographic dimension is relevant, researchers have two different alternatives to define the basic territorial units that will be used in the study: To use geographical units designed following normative criteria or to apply an analytical criteria to identify these units.

"Normative regions are the expression of a political will; their limits are fixed according to the tasks allocated to the territorial communities, to the sizes of population necessary to carry out these tasks efficiently and economically, or according to historical, cultural and other factors. Whereas analytical (or functional) regions are defined according to analytical requirements: functional regions are formed by zones grouped together using geographical criteria (e.g.,

altitude or type of soil) or/and using socio-economic criteria (e.g., homogeneity, complementarity or polarity of regional economies)" (Eurostat, 2004).

The majority of empirical studies tend to use geographical units based on normative criteria for several reasons: this type of units are officially established, they have been traditionally used in other studies, its use makes comparison of results easier and can be less criticized. But at the same time, in those studies using this type of units an "Achilles' heel" can exist if they are very restrictive or inappropriate for the considered problem. For example, if we are analysing phenomena as regional effects of monetary and fiscal policy, how will the results be affected if the aggregated areas in each region are heterogeneous? can those results change if the areas are redefined in a way that each region contains similar areas?.

The above mentioned situation could be improved through the use of automated regionalisation tools specialized on design geographical units based on analytical criteria. In this context, the design of analytical geographical units should consider the following three fundamental aspects:

i. *Geographical contiguity*: The aggregation of areas (small spatial units) into regions such that the areas assigned to a region must be internally connected or contiguous.

ii. *Equality*: In some cases, it is important that designed regions are "equal" in terms of some variable (for example population, size, presence of infrastructures, etc).

iii. *Interaction between areas*: Some variables do not exactly define geographical characteristics that can be used to aggregate the different areas, but perhaps they describe some kind of interactions among them (for example, distance, time, number or trips between areas, etc). These variables can also be used as interaction variables using some dissimilarity measure between areas in terms of socio-economic characteristics. The objective in this kind of regionalisation process is that areas belonging to the same region are as homogeneous as possible with respect to the specified attribute(s).

The two most used methodological strategies to design analytical geographical units consists in, first, to apply conventional clustering algorithms and, second, to use additional instruments to control for the continuity restriction. In this paper, we will use both strategies, which are, next, briefly described:

a) Two stages strategy:

In order to apply conventional clustering algorithms, it is necessary to split the regionalisation process into two stages. The first stage consists in applying a conventional clustering model without taking into account the contiguity constraint. In the second stage, the clusters are revised in terms of geographical contiguity. With this methodology, if the areas included in the same cluster are geographically disconnected those areas are defined as different regions (Ohsumi, 1984).

Among the advantages of this methodology, Openshaw and Wymer (1995) highlighted that the homogeneity of the defined regions is guaranteed by the first stage. Moreover, this methodology can also be useful as a way to obtain evidence of spatial dependence among the elements. However, taking into account the objectives of the regionalisation process, the fact that the number of groups depends on the degree of spatial dependence and not on the researcher criteria can be an important problem.

Two conventional clustering algorithms can be used in this context: hierarchical or partitional. In this paper, we apply the K-means clustering procedure, which belongs to partitional clustering category[3].

The K-means clustering is an iterative technique that consists in selecting from elements to be grouped, a predetermined number of *k* elements that will act as centroids (the same number as groups to be formed). Then, each of the other elements is assigned to the closest centroid.

The aggregation process is based on minimizing some measure of dissimilarity among elements to aggregate in each cluster. This dissimilarity measure is usually calculated as the squared Euclidean distance from the centroid of the cluster[4].

$$\sum_{m \in c} \sum_{i=1}^{N} \left(X_{im} - \overline{X}_{ic} \right)^2 \qquad (1)$$

Where X_{im} denotes the value of variable *i* (*i*=1..*N*) for observation *m* (*m*=1..*M*), and $\overline{X}_{ic}$ is the centroid of the cluster *c* to which observation *m* is assigned or the average X_i for all the observations in cluster *c*.

K-means algorithm is based on an iterative process where initial centroids are explicitly or randomly assigned and the other elements are assigned to the nearest centroid. After this initial assignation, initial centroids are reassigned in order to minimize the squared Euclidean distance. The iterative process is terminated if there is not any change that would improve the actual solution.

[3] Hierarchical algorithms are usually applied when the researcher is interested in obtain a hierarchical and nested classification (for every scale levels). The main disadvantage of using hierarchical clustering algorithms is the high probability of obtaining local optimum due to the fact that once two elements have been grouped in an aggregation level, they would not return to be evaluated independently in higher aggregation leves (Semple and Green, 1984).

[4] A detailed summary of these aggregation methodologies can be found in Gordon (1999) and for the case of constrained clustering in Fisher (1980), Murtagh (1985) and Gordon (1996).

It is important to note that the final solutions obtained by applying K-means algorithm depend on the starting point (the initial centroids designation). This fact makes quite difficult to obtain a global optimum solution.

Finally, when K-means algorithm is applied in a two stages regionalisation process, it will be possible that the required number of regions to design will be not necessarily equal to the value given to parameter *k* as areas belonging to the same cluster have to be counted as different regions if they are not contiguous. So, different proofs have to be done with different values of *k* (lower than the number of desired regions), until contiguous regions are obtained.

b) Additional instruments to control for the continuity restriction:

It is possible to control the geographical contiguity constraint using additional instruments as the contact matrix or its corresponding contiguity graph. Those elements are used to adapting conventional clustering algorithms, hierarchical or partitioning, with the objective of respecting the continuity constraint.

The partitioning algorithm used in this paper applies a recently linear optimisation model proposed by Duque, Ramos and Suriñach (2004). The heterogeneity measure used in this model consists in the sum of the dissimilarities between areas in each region. Following Gordon (1999), the heterogeneity measure for region *r*, C_r can be calculated as follows:

$$H(C_r) \equiv \sum_{\{i,j \in C_r | i<j\}} d_{ij} \qquad (2)$$

Taking this into account, the problem of obtaining *r* homogeneous classes (regions) can be understood as the minimisation of the sum of the heterogeneity measures of each class (region) *r*:

$$P(H,\Sigma) \equiv \sum_{r=1}^{c} H(C_r) \qquad (3)$$

The objective function of the optimisation model looks for the minimisation of the total heterogeneity, measured as the sum of the elements of the upper triangular matrix (D_{ij}) of dissimilarity relationships between areas belonging to the same region (the elements defined by the binary matrix T_{ij}).

$$\text{Objective function}: Min \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} \cdot T_{ij} \qquad (4)$$

Where $D_{i,j}$ is the value of the dissimilarity relationships between areas *i* and *j*, with *i*<*j*; and T_{ij} is a binary matrix where elements *ij* are equal to 1 if areas *i* and *j* belong to the same region and 0 otherwise.

The main characteristics of this optimisation model are the following:

i. Automated regionalisation model that allow to design a given number of homogeneous geographical units from aggregated small areas subject to contiguity requirements.

ii. To formulate the regionalisation problem as a lineal optimisation problem ensures the possibility of finding the global optimum among all feasible solutions.

iii. More coherent solutions can be easily obtained introducing additional constraints related to other specific requirements that are relevant for the regionalisation process.

iv. With this model a region consist of two or more contiguous areas, it implies that any region can be formed by a unique area[5].

In order to apply this model in bigger regionalisation processes, the model is incorporated into an algorithm called RASS (Regionalisation Algorithm with Selective Search) proposed by Duque, Ramos and Suriñach (2004). The most relevant characteristic of this new algorithm is related to the fact that the way it operates is inspired in the own characteristics of regionalisation processes, where available information about the relationships between areas can play a crucial role in directing the searching process in a more selective and efficient way (i.e. less random). In fact, the RASS incorporates inside its algorithm the optimisation model we present above in order to achieve local improvements in the objective function. These improvements can generate significant changes in regional configurations; changes that would be very difficult to obtain using other iterative methods.

3. Normative vs. analytical regions: The case of regional unemployment in Spain

There are many economic variables whose analysis at a nationwide aggregation level is not representative as a consequence of important regional disparities. These regional disparities make necessary to complement the aggregated analysis with applied research at a lower aggregation level in order to have a better knowledge of the studied phenomenon. A clear example of this case can be found when analysing the unemployment rate. Previous studies have demonstrated that Spanish unemployment rate presents important disparities (Alonso and Izquierdo, 1999), accompanied of spatial dependence (López-Bazo

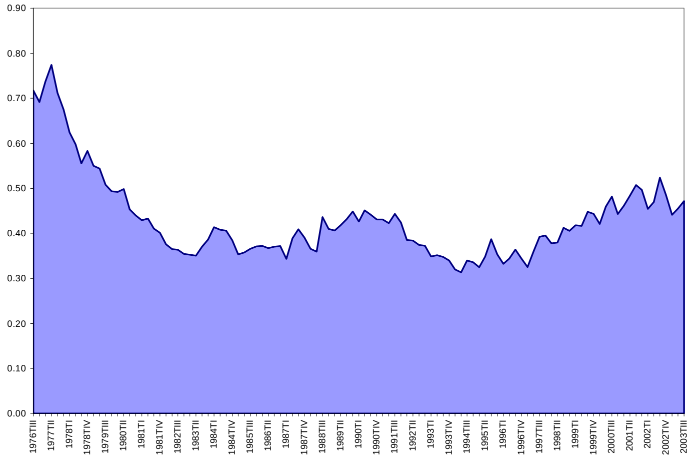
[5] As Crone (2003) highlights, this is one of the conditions followed by the Bureau of Economic Analysis (BEA) for the regionalisation of the United States of America.

et al. 2002) at the provincial aggregation level (NUTS I). In fact. these two elements, disparity and spatial dependence, make of this variable a good candidate to make regionalisation experiments that allow to analyse the differences that can be generated between the normative and analytical geographical divisions. The analysis in this section focuses on quarterly provincial unemployment rates in peninsular Spain from the third quarter of 1976 to the third quarter of 2003.

First of all, some descriptive will be presented in order to confirm the existence of spatial differences and dependence.

Regarding spatial disparity, figure 1 shows the variation coefficient of NUTS III unemployment rates during the considered period. As it can be seen, throughout the analysed period, we observe an important dispersion of the unemployment rate between Spanish provinces with an average value for the whole period of 43.03%. This dispersion was considerably higher during the second half of the 70's. These disparities are obvious if we take into account that the average difference between maximum and minimum rates during the considered period was 25.59.

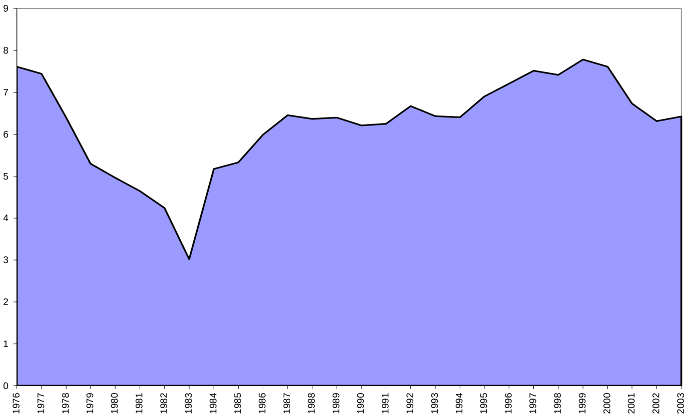
Figure 1. Variation coefficient for the unemployment rate at NUTS III level



Source: Own elaboration

Regarding spatial dependence, we have calculated the Moran's *I* statistic (Moran, 1948)[6] of first-order spatial autocorrelation. The values for the standardized Moran's *I* Z(I), which follows an asymptotical normal standard distribution, for the provincial unemployment rate during the considered period is shown in figure 2. As it can be seen, all Z-values are greater than 2 indicating that the null hypothesis of a random distribution of the variable throughout the territory (non spatial autocorrelation) should be rejected.

Figure 2. Z-Moran statistic for the unemployment rate at NUTS III level[7]



Source: Own elaboration

After the above descriptive analysis, the possibility of carrying out a regionalisation process is clearly justified: The existence of spatial differences gives rise to the creation of groups, whereas the spatial dependence justifies the imposition of geographical contiguity of these groups.

So, with the objective to compare the results obtained when making an analytical regionalisation process with the territorial division NUTS, which have been established according to normative criteria, we will design regions based on the behaviour of the provincial unemployment such that provinces belonging

[6] More information about this statistic is provided in annex 1.

[7] The values of this statistic have been calculated using the "SPSS Macro to calculate Global/Local Moran's I" by M. Tieseldorf. http://128.146.194.110/StatsVoyage/Geog883.01/SPSS%20Moran%20Macro.htm.

to the same region would be as homogeneous as possible in terms of this variable.

In order to facilitate the comparison with NUTS division, two scale levels have been established. The first one forms 15 regions to be compared to the 15 regions in which the peninsular Spain is divided at the NUTS II level, while the second scale has been set to 6 in order to be compared with NUTS I division.

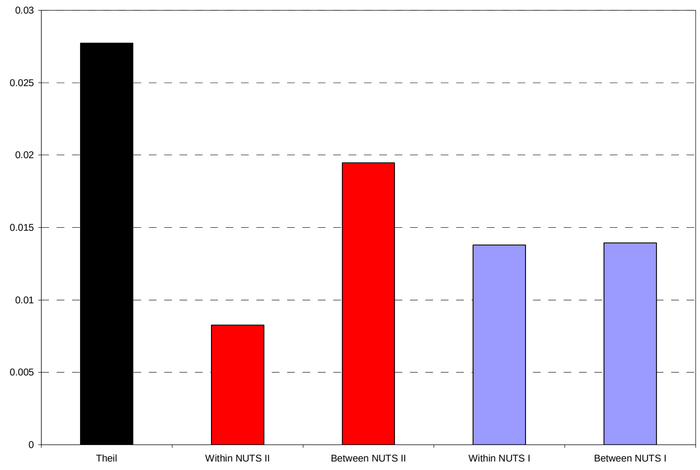
One way of comparing the homogeneity[8] of the different territorial divisions consists in calculating the Theil's inequality index (Theil, 1967). One advantage of this index in this context is that it permits the decomposition of its value into two components a within and a between component. The aim of analytical regionalisation procedures should be to minimise within inequalities[9] and maximise between inequalities.

Figure 3 shows the total value of the Theil's inequality index and the value of the within and between components when average unemployment rates of Spanish provinces (NUTS III) are aggregated into NUTS II and NUTS I regions. The most relevant result from this figure is that the level of "internal" homogeneity (the within component) is very high (in relative terms) for both scale levels, but in particular at the NUTS I level.

[8] Conceição *et al* (2000) apply the Theil Index to data on wages and employment by industrial classification to measure the evolution of wage inequality through time.

[9] See annex 2 for more information on this statistic.

Figure 3. Decomposition of the Theil index for the unemployment rate for NUTS III regions into NUTS II and NUTS I regions



Source: Own elaboration

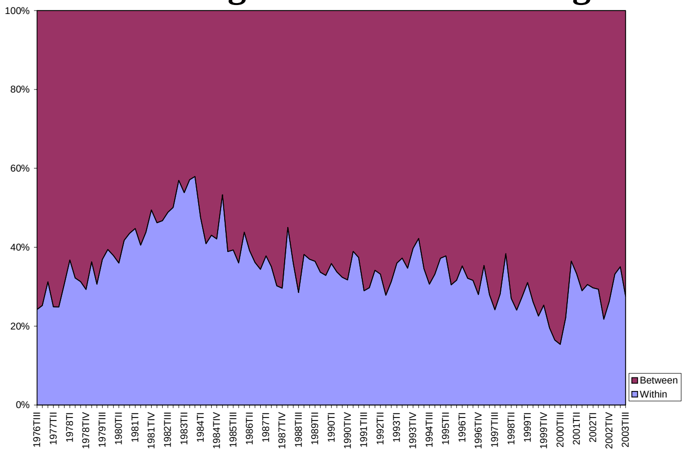
An important goal when normative regions (NUTS) are designed is that those regions should minimise the impact of the (inevitable) process of continuous change in regional structures. But, regarding to the provincial unemployment rate, are the NUTS regions representative of the behaviour of regional unemployment during the whole period?. Figures 4 and 5 show the relative decomposition of the Theil's inequality index along the analysed period. For both, NUTS II (figure 4) and NUTS I (figure 5) it can be seen that within inequality depicts an irregular behaviour, showing the greater dispersion at the beginning of the eighties. The highest homogeneity level is reached during 2000. It is also important to note that the proportion of within inequality in NUTS I is strongly higher that in NUTS II, in part, because at a smaller scaling level (from 15 to 6 regions) the differences within the groups tend to increase. This aggregation impact becomes worse due to nested aggregation of NUTUS II to obtain NUTS I[10]

Can an analytical regionalisation process improve the results obtained for normative regions? In order to answer this question, two stages and optimisation model regionalisation algorithms have been applied.

[10] That disadvantage was commented above, in section 2, when hierarchical aggregation was introduced.

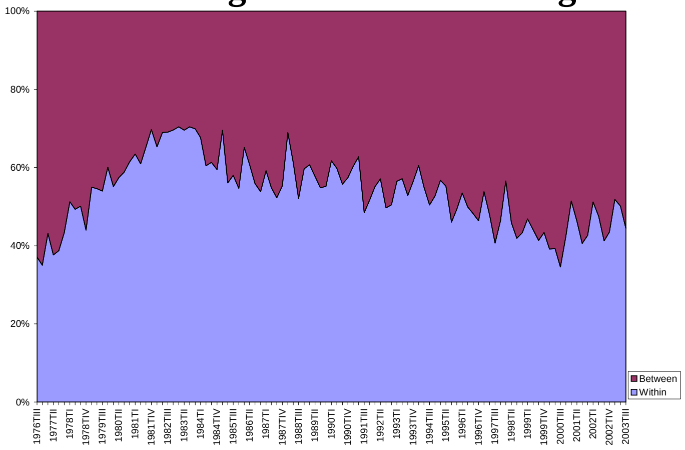
The K-means algorithm have been applied to the unemployment rates to group the 47 contiguous provinces into 15 and 6 regions, These results will be compared with the normative regions (NUTS II and NUTS I) presented above. The same process will also be done by applying the RASS algorithm. And, last, a comparison between K-means and RASS is done.

Figure 4. Decomposition of the Theil index for the unemployment rate for NUTS III regions into NUTS II region



Source: Own elaboration

Figure 5. Decomposition of the Theil index for the unemployment rate for NUTS III regions into NUTS I regions



Source: Own elaboration

It is important to note that dissimilarities between provinces calculated by K-means and RASS algorithms takes into account the whole period (from 1976-QIII to 2003-QIII). This strategy provides to the regionalisation process a dynamic component with the aim of designing temporally representatives regions. The use of Euclidean distances (squared in K-means) allows taking into account both, the direction and magnitude differences between the values of unemployment rates of the different areas.

Figure 6 shows a comparison between normative and analytical regions using K-means. The values below the provincial code indicate the deviation from the arithmetic average (unweighted) of the unemployment rate of the region which it belongs[11]. It is expected that if regions are homogeneous, then the provincial unemployment rate should be near to the regional one.

For NUTS II (left side map) the maximum deviations are located in Barcelona (number 8 in the map) with 6.06% over the regional average, and Almería (4), with 7.83% under the regional average. It is worth mentioning that the range is 13.88, a value that indicates important differences in the unemployment rate between provinces belonging to the same region.

With respect to analytical regions obtained by K-means (right side map), the deviations are lower than in the NUTS II case: the maximum value is now 2.16% (Valladolid - 44) and the minimum value is -2.22% (Lugo - 27). In this case, the range is 4.38, which is substantially lower than before.

Once 15 analytical aggregations have been designed in order to be compared to NUTS II, the unemployment rate has been re-calculated for each one of the 15 regions. The new series have been used to aggregate those 15 regions into 6 analytical regions. This methodology ensures that the obtained aggregation are nested into the previous one in a way that permits comparison to NUTS I. It is important to note that when K-means cluster was applied, it was impossible to obtain six regions, because we had to fix the number of cluster

[11] As the simple average was calculated, for each region, the sum of provincial deviations is equal to zero.

regions to three to obtain contiguous regions, and then the number of contiguous regions was seven[12].

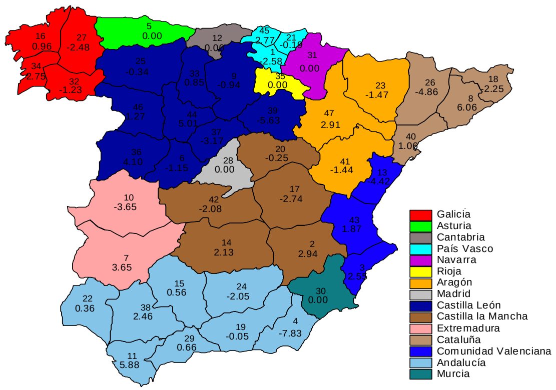
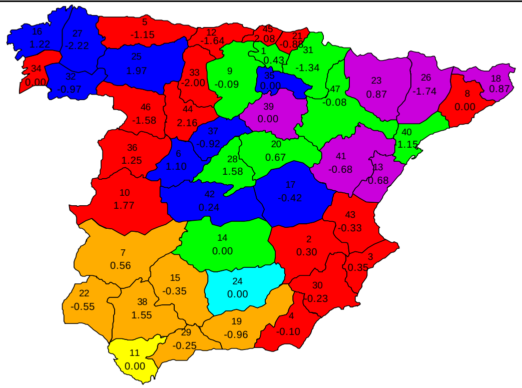
Figure 7 shows normative regions (left side map) that correspond to NUTS I aggregation level, and analytical regions (right side map). Again, lower deviations are obtained for the analytical regions. For NUTS I regions, the maximum value of the deviation is 10.86% in Badajoz (7) and the minimum is –7.08% in Murcia (30). For analytical regions, the values are 4.72% (Cadiz - 11) and –3.53% (Navarra - 31). Now, the range has decreased from 17.93 to 8.25.

For a more detailed analysis, in terms of the homogeneity reached by using analytical regionalisation with K-means algorithm, the Theil's inequality index was again calculated. The results in figure 8 show an important improvement in terms of within/between inequality. In both cases, CLUSTER II and CLUSTER I aggregation levels, inequality within regions represents only a 4.68% and a 11.98% of total inequality between provinces. This implies that analytical regions are much more homogeneous than normative ones in terms of average unemployment rates.

Another relevant result is obtained when the Theil's inequality index is calculated for each quarter for the different aggregation levels (figures 9 and 10). As it can be seen, within inequality is more constant for analytical regions than for normative regions.

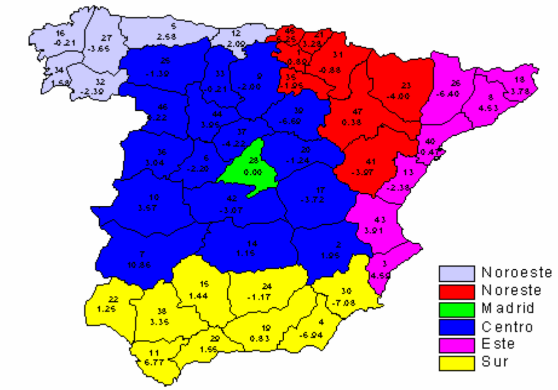
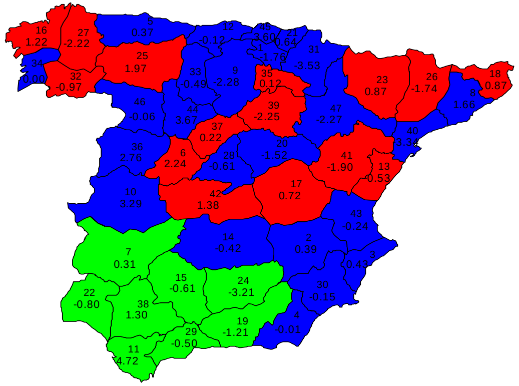
[12] If the value of the cluster regions was set to two, then only two contiguous would have been obtained.

Figure 6. Comparison between administrative (NUTS II) and economic regions using the K-means cluster

NUTS II	Cluster (K-means) II
	

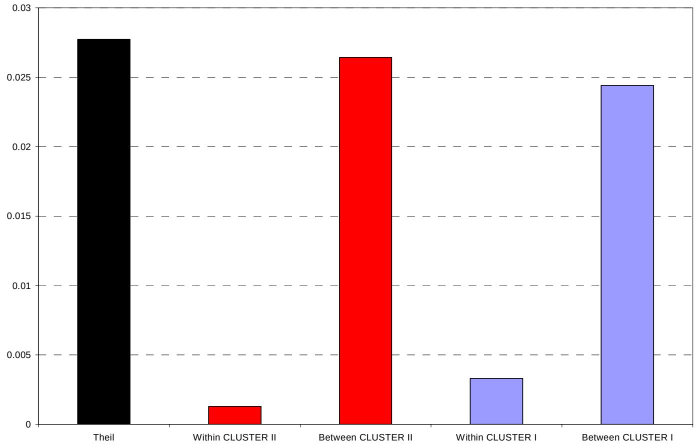
Source: Own elaboration

Figure 7. Comparison between administrative (NUTS I) and economic regions using the K-means cluster

NUTS I	K-means I
	

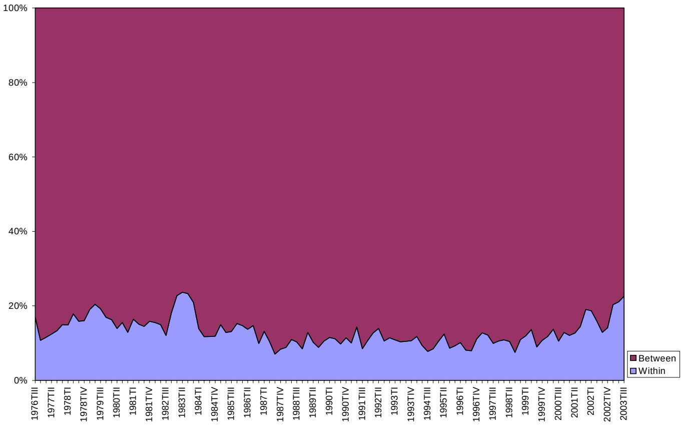
Source: Own elaboration

Figure 8. Decomposition of the Theil index for the unemployment rate for NUTS III regions into Cluster II and Cluster I regions



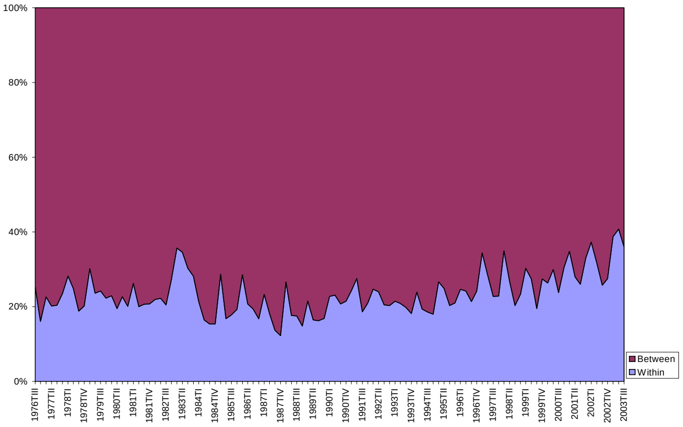
Source: Own elaboration

Figure 9. Decomposition of the Theil index for the unemployment rate for NUTS III regions into Cluster II regions



Source: Own elaboration

Figure 10. Decomposition of the Theil index for the unemployment rate for NUTS III regions into Cluster I regions

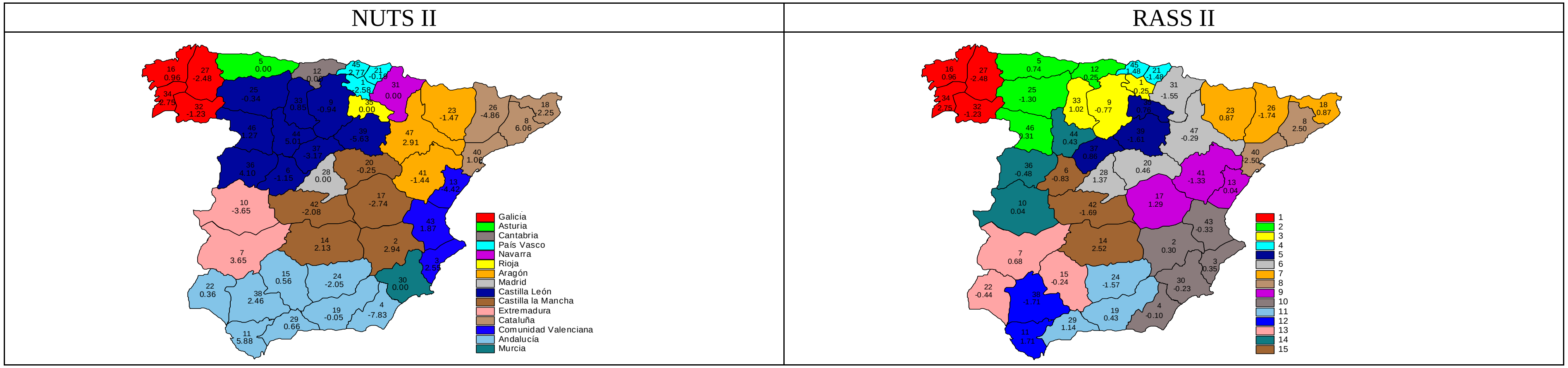


Source: Own elaboration

The second analytical regionalisation procedure applied in this paper is the RASS algorithm. Figures 11 and 12 show the analytical regions obtained applying RASS and the normative regions (NUTS) for the two considered aggregation levels. In both levels, the average unemployment rates show lower deviations with respect regional averages when using RASS. In RASS II, Pontevedra (34) and Tarragona (40) present the higher deviations (2.75%) and the lower (-2.50%). In RASS I aggregation, the extreme deviations are located in Barcelona (8) and Lleida (26) with a deviation from regional averages of 6.51% and -4.42%, respectively. In both cases, the ranges are considerably lower in RASS regions than in normative regions, as in the K-means case.

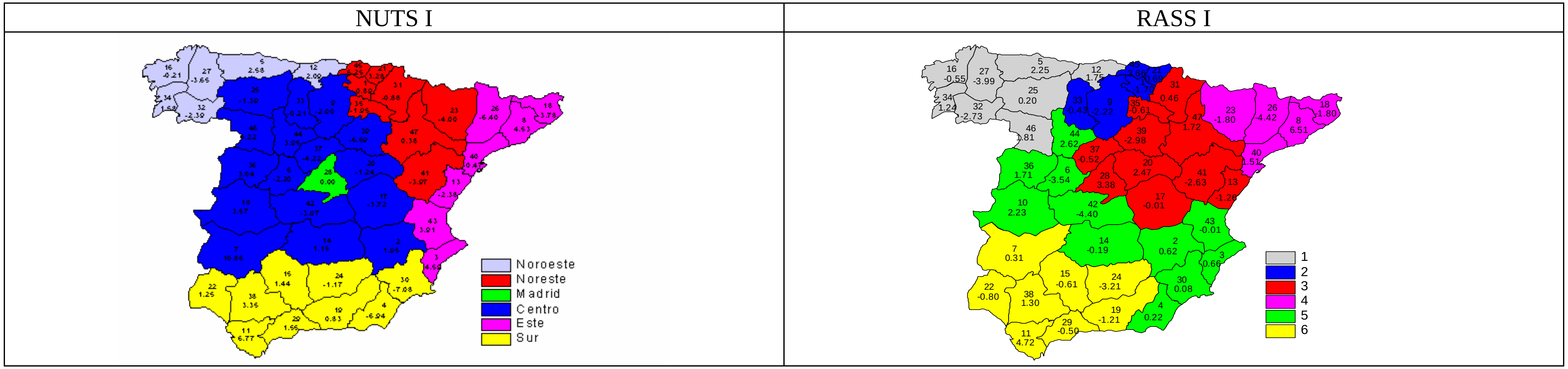
The values of the Theil's inequality index (figure 13), calculated for RASS II and RASS I regions using the average unemployment rates, show that the inequality within regions is strongly reduced to a 6.54% and a 21.64% of the total inequality. This fact implies that, again, analytical regions using RASS are much more homogeneous that normative ones in terms of average unemployment rates. In RASS II, the within inequality remains relatively constant along the analysed period (figure 14), but for RASS I (figure 15) the within inequality is especially higher between 1976 and 1984.

Figure 11. Comparison between administrative (NUTS II) and economic regions using the RASS procedure



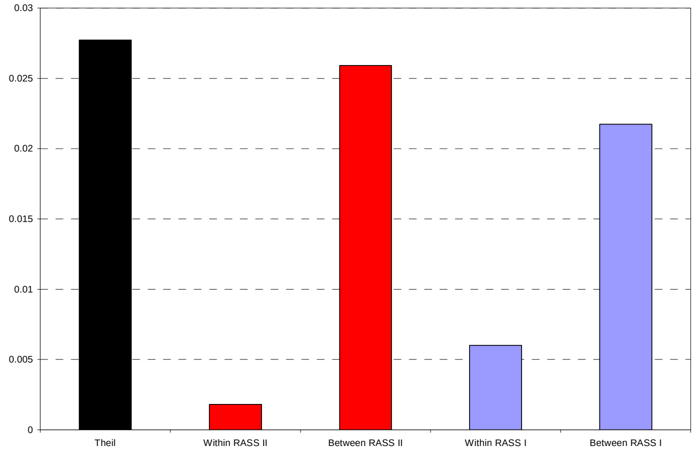
Source: Own elaboration

Figure 12. Comparison between administrative (NUTS I) and economic regions using the RASS procedure



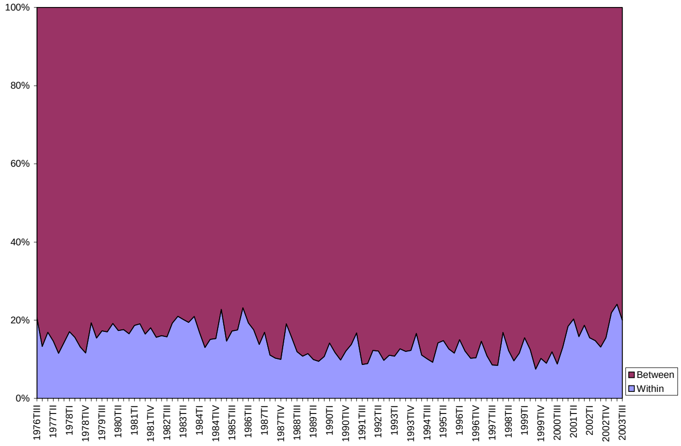
Source: Own elaboration

Figure 13. Decomposition of the Theil index for the unemployment rate for NUTS III regions into RASS II and RASS I regions



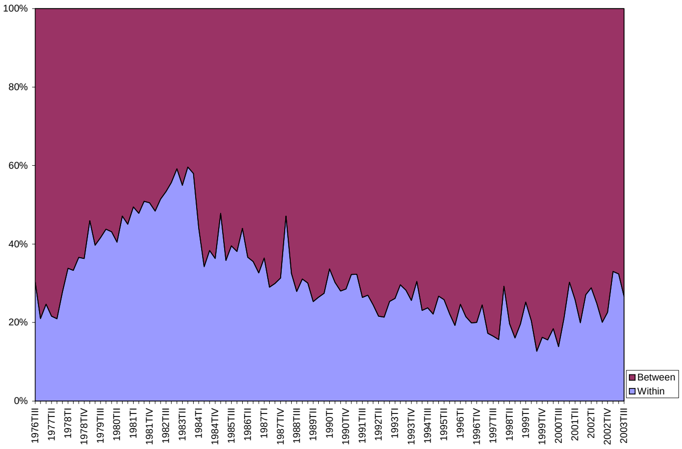
Source: Own elaboration

Figure 14. Decomposition of the Theil index for the unemployment rate for NUTS III regions into RASS II regions



Source: Own elaboration

Figure 15. Decomposition of the Theil index for the unemployment rate for NUTS III regions into RASS I regions



Source: Own elaboration

Table 1 summarises the basic descriptive statistics commented above. In fact, these statistics establish the basis for a comparison between the different regionalisation procedures applied. This comparison has been divided into different regionalisation characteristics: Homogeneity, regional shape, control level and flexibility. In each category the main advantages or disadvantages of each analytical method will be mentioned.

Homogeneity: Both analytical regionalisation methods improve strongly the intra-regional homogeneity along the whole period. For both aggregation levels (II and I), Clustering method (using K-means algorithm) obtains lower values of within regional dispersion (see table 1).

Table 1. Descriptive statistics for the different regional classifications

	NUTS II	RASS II	CLUSTER II	NUTS I	RASS I	CLUSTER I
Maximum	6.06	2.75	2.16	10.86	6.51	4.72
Minimum	-7.83	-2.50	-2.22	-7.08	-4.42	-3.53
Range	13.88	5.25	4.38	17.93	10.92	8.25
Standard deviation	1.90	0.74	0.69	2.30	1.49	1.21

Source: Own elaboration

Regional shape: With respect to the final regional shape obtained with analytical regionalisation methods, two stages strategy tends to design strongly irregular region shapes compared with the RASS strategy. If more compact regions are desired, the geographical coordinates of the points representing the areas to be aggregated could be included in the calculation of dissimilarities between areas (Perruchet, 1983, Webster and Burrough, 1972). However, the weight that has to be assigned to this new component inside the dissimilarities calculation can only be based on subjective criteria[13]. Also, with the two stages strategy, the number of provinces grouped in each region shows big differences: in Cluster II there are seven

[13] For a more detailed discussion about this problem, see Wise, Haining and Ma, 1997.

regions formed by one province, while there are regions formed by nine provinces. The same happens in Cluster I, since the number of provinces assigned to a region takes values between one and seventeen. On the other hand, RASS methodology forms more balanced regions: at RASS II, the number of provinces by regions varies between two and four, and, it varies between five and eleven at RASS I.

Control level: One of the main disadvantages in two stages strategy is that the researcher does not have total control with regard to the number of regions to be designed. It can be seen in Cluster I, where it was impossible to obtain six regions. This kind of problem does not exist in RASS algorithm because the number of regions to be designed is a given parameter in the model.

Flexibility: This characteristic is very important when the researcher wants to introduce additional constraints in the regionalisation process. In this case, the RASS algorithm has an important advantage compared with the K-means algorithm. In the RASS method, additional constrains can be imposed by introducing them explicitly as additional constraints in the model or by formulating a multiobjective function. Those constrains could be related to aspects such as area characteristics or with areas relationships.

4. Final remarks

Two different regionalisation processes were applied in order to design analytical regions that are homogeneous in terms of the interest variable: one based in the application of the K-means algorithm and a second one based on mathematical programming (RASS algorithm).

Both models were applied in the context of provincial unemployment rates in Spain in order to compare normative with the obtained analytical regions. The results have shown that more homogeneous regions were designed when applying both analytical regionalisation tools. Two other obtained interesting results are related with the fact that analytical regions were also more stable along time and with the effects of scale in the regionalisation process.

5. References

- Albert, J. M., Mateu, J. and Orts, V. (2003), *Concentración versus dispersion: Un análisis especial de la localización de la actividad económica en la U.E.*, mimeo.
- Alonso, J. and Izquierdo, M. (1999), 'Disparidades regionales en el empleo y el desempleo', *Papeles de Economía Española*, 80, 79-99.
- Conceição, P., Galbraith, J. K. and Bradford, P. (2000), 'The Theil Index in Sequences of Nested and Hierarchic Grouping Structures: Implications for the Measurement of Inequality through Time with Data Aggregated at Different Levels of Industrial Classification', UTIP Working Paper Number 15.
- Crone, T. M. (2003), 'An alternative definition of economic regions in the U.S. based on similarities in State business cycles', Federal Reserve Bank of Philadelphia, Working Paper 03-23.
- Duque, J.C., Ramos, R. and Suriñach, J. (2004), 'Design of Homogenous Territorial Units: A Methodological Proposal', Documents de Treball de la Divisió de Ciències Jurídiques, Econòmiques i Socials. Universitat de Barcelona. forthcoming.
- Eurostat, 2004. 'Nomenclature of territorial units for statistics – NUTS. Statistical Regions of Europe'. http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html. (01/03/04).
- Fisher, M. M. (1980), 'Regional taxonomy', *Regional Science and Urban Economics*, 10, 503-37.
- Gordon, A. D. (1996), 'A survey of constrained classification', *Computational Statistics & Data Analysis*, 21, 17-29.
- Gordon, A. D. (1999), Classification (second edition ed.). Boca Raton [etc.].

- López-Bazo, E., Vaya, E., Mora, A. and Suriñach, J. (1999), ‘Regional Economic Dynamics and Convergence in the European Union’, *Annals of Regional Science*, 33, 343-370.
- López-Bazo, E., del Barrio, T. and Artís, M. (2002), ‘The regional distribution of Spanish unemployment:: A spatial analysis’, *Papers in Regional Science*, 81, 365-389.
- Moran, P. (1948), ‘The interpretation of statistical maps’, *Journal of the Royal Statistical Society B*, 10, 243-251.
- Murtagh, F. (1985), ‘A survey of Algorithms for Contiguity-constrained Clustering and Related Problems’, *The Computer Journal*, 28 (1), 82-88.
- Ohsumi, N. (1984), ‘Practical techniques for areal clustering’, in *Data analysis and informatics*, Vol III, E. Diday, M. Jambu, L. Lebart, J. Pagès and R. Tomassone, (eds.) Vol. III. North-Holland, Amsterdam, pp 247-58.
- Openshaw, S. (1984), ‘The modifiable areal unit problem’, *Concepts and Techniques in Modern Geography*, 38 (GeoAbstracts, Norwich).
- Openshaw, S. and Wymer, C. (1995), ‘Classifying and regionalizing census data’, in *Census Users Handbook*, S. Openshaw, (eds.). Cambridge, UK: Geo Information International, pp 239-70.
- Perruchet, C. (1983), ‘Constrained agglomerative hierarchical classification’, *Pattern Recognition*, 16, 213-17.
- Semple, R. K. and Green, M. B. (1984), ‘Classification in human geography’, in *Spatial statistics and models*, G. L. Gaile and C. J. Wilmott, (eds.). Reidel, Dordrecht, pp 55-79.
- Theil, H. (1967). Economics and Information Theory. Chicago: Rand McNally and Company.
- Webster, R. and Burrough, P. A. (1972), ‘Computer-based soil mapping of small areas from sample data II. Classification smoothing’, *Journal of Soil Science*, 23, 222-34.
- Wise, S. M., Haining, R. P. and Ma, J. (1997), ‘Regionalization Tools for Exploratory Spatial Analysis of Health Data’, in *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling, and computational intelligence*, Manfred M. Fisher and Arthur Gentis, (eds.). Berlin [etc.]: Springer, pp 83-100.

6. Annexes

Annex 1. Moran's *I*:

$$I = \frac{\sum_{ij}^{N} w_{ij}\left(x_i - \bar{x}\right)\cdot\left(x_j - \bar{x}\right)}{\left(x_i - \bar{x}\right)^2} \qquad\qquad i \neq j$$

For each quarter, x_i and x_j are unemployment rates in provinces *i* and *j*,. $\bar{x}$ is the average of the unemployment rate in the sample of provinces; and w_{ij} is the *ij* element of a row-standarized matrix of weights (we used the binary contact matrix, it is a binary matrix with elements w_{ij}, where w_{ij} takes value 1 if areas *i* and *j* share a border; and 0 otherwise)

Annex 2. Theil Index:

$$T = \sum_{p=1}^{n} \frac{u_p}{U} \log\left[\frac{\left(\frac{u_p}{U}\right)}{\left(\frac{1}{n}\right)}\right]$$

Where n is the number of provinces (47), u_p is the provincial unemployment rate indexed by p, and U representing the Spanish unemployment rate $U = \sum_{p=1}^{n} u_p$

Overall inequality can be completely and perfectly decomposed into a between-group component $T_g^{'}$, and a within-group component (T_g^W). Thus: $T = T_g^{'} + T_g^W$. With $T_g^{'} = \sum_{i=1}^{m} \frac{U_i}{U} \log\left[\frac{\frac{U_i}{U}}{\frac{n_i}{n}}\right]$ where i indexes regions, with n_i representing the number of provinces in group i, and U_i the unemployment rate in region i., and $T_g^W = \sum_{t=1}^{m} \frac{U_i}{U} \sum_{p=1}^{n_i} \frac{u_{ip}}{U_i} \log\left[\frac{\left(\frac{u_{ip}}{U_i}\right)}{\left(\frac{1}{n_i}\right)}\right]$, where each provincial unemployment rate is indexed by two subscripts: i for the unique region to which the province belongs, and subscript p, where, in each region, p goes from 1 to n_i.

Annex 3. Regional configurations

Table A.1. NUTS Classification for the Spanish regions

NUTS I	NUTS II	NUTS III	CODE
NOROESTE	GALICIA	Coruña (A)	16
		Lugo	27
		Orense	32
		Pontevedra	34
	ASTURIA	Asturias	5
	CANTABRIA	Cantabria	12
NORESTE	PAIS VASCO	Álava	1
		Guipúzcoa	21
		Vizcaya	45
	NAVARRA	Navarra	31
	RIOJA	Rioja (La)	35
	ARAGON	Huesca	23
		Teruel	41
		Zaragoza	47
MADRID	MADRID	Madrid	28
CENTRO	CASTILLA LEON	Ávila	6
		Burgos	9
		León	25
		Palencia	33
		Salamanca	36
		Segovia	37
		Soria	39
		Valladolid	44
		Zamora	46
	CASTILLA LA MANCHA	Albacete	2
		Ciudad Real	14
		Cuenca	17
		Guadalajara	20
		Toledo	42
	EXTREMADURA	Badajoz	7
		Cáceres	10
ESTE	CATALUÑA	Barcelona	8
		Girona	18
		Lleida	26
		Tarragona	40
	COMUNIDAD VALENCIANA	Alicante	3
		Castellón de la Plana	13
		Valencia	43
SUR	ANDALUCIA	Almería	4
		Cádiz	11
		Córdoba	15
		Granada	19
		Huelva	22
		Jaén	24
		Málaga	29
		Sevilla	38
	MURCIA	Murcia	30

Source: Eurostat

Table A.2. Detailed results of the regionalisation process using the K-means cluster procedure

Cluster I	Cluster II	NUTS III	CODE
1	1	Pontevedra	34
2	2	Coruña (A)	16
		León	25
		Lugo	27
		Orense	32
3	3	Asturias	5
		Cáceres	10
		Cantabria	12
		Guipúzcoa	21
		Palencia	33
		Salamanca	36
		Valladolid	44
		Vizcaya	45
		Zamora	46
	4	Álava	1
		Burgos	9
		Guadalajara	20
		Madrid	28
		Navarra	31
		Tarragona	40
		Zaragoza	47
	8	Barcelona	8
4	7	Girona	18
		Huesca	23
		Lleida	26
5	5	Rioja (La)	35
	6	Soria	39
	9	Castellón de la Plana	13
		Teruel	41
	15	Ávila	6
		Cuenca	17
		Segovia	37
		Toledo	42
6	10	Albacete	2
		Alicante	3
		Almería	4
		Murcia	30
		Valencia	43
	14	Ciudad Real	14
7	11	Badajoz	7
		Córdoba	15
		Granada	19
		Huelva	22
		Málaga	29
		Sevilla	38
	12	Cádiz	11
	13	Jaén	24

Source: Own elaboration

Table A.3. Detailed results of the regionalisation process using the RASS procedure

RASS I	RASS II	NUTS III	CODE
1	1	Coruña (A)	16
		Lugo	27
		Orense	32
		Pontevedra	34
	2	Asturias	5
		Cantabria	12
		León	25
		Zamora	46
2	3	Álava	1
		Burgos	9
		Palencia	33
	4	Guipúzcoa	21
		Vizcaya	45
3	5	Rioja (La)	35
		Segovia	37
		Soria	39
	6	Guadalajara	20
		Madrid	28
		Navarra	31
		Zaragoza	47
	9	Castellón de la Plana	13
		Cuenca	17
		Teruel	41
4	7	Girona	18
		Huesca	23
		Lleida	26
	8	Barcelona	8
		Tarragona	40
5	10	Albacete	2
		Alicante	3
		Almería	4
		Murcia	30
		Valencia	43
	14	Cáceres	10
		Salamanca	36
		Valladolid	44
	15	Ávila	6
		Ciudad Real	14
		Toledo	42
6	11	Granada	19
		Jaén	24
		Málaga	29
	12	Cádiz	11
		Sevilla	38
	13	Badajoz	7
		Córdoba	15
		Huelva	22

Source: Own elaboration