UNIVERSITAT DE BARCELONA

Design of Homogenous Territorial Units. A Methodological Proposal and Applications

Juan Carlos Duque Cardona

Departamento de Econometría Estadística y Economía Española

UNIVERSIDAD DE BARCELONA

DESIGN OF HOMOGENOUS TERRITORIAL UNITS.

A Methodological Proposal and Applications.

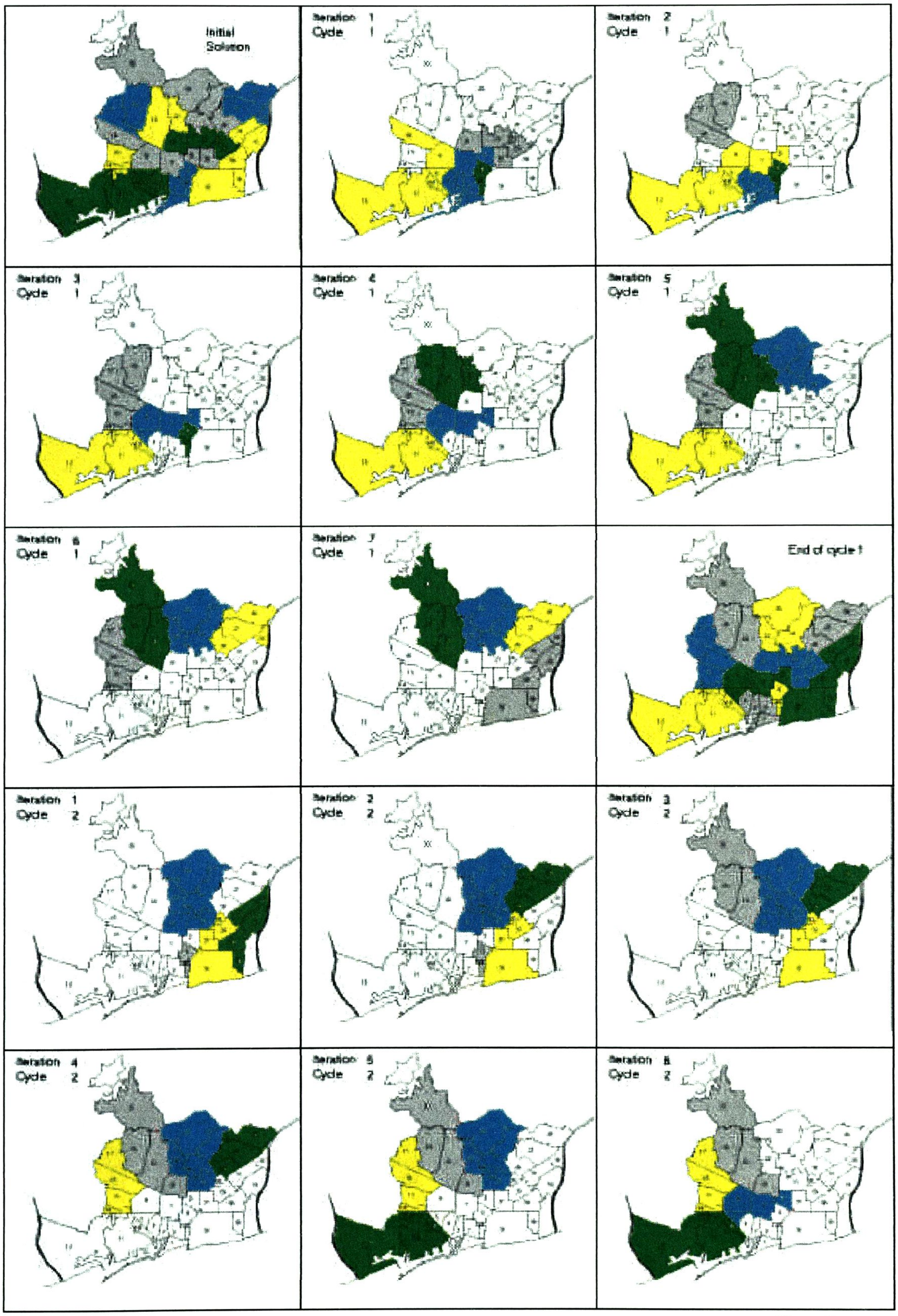
Juan Carlos Duque Cardona

MAYO 2004

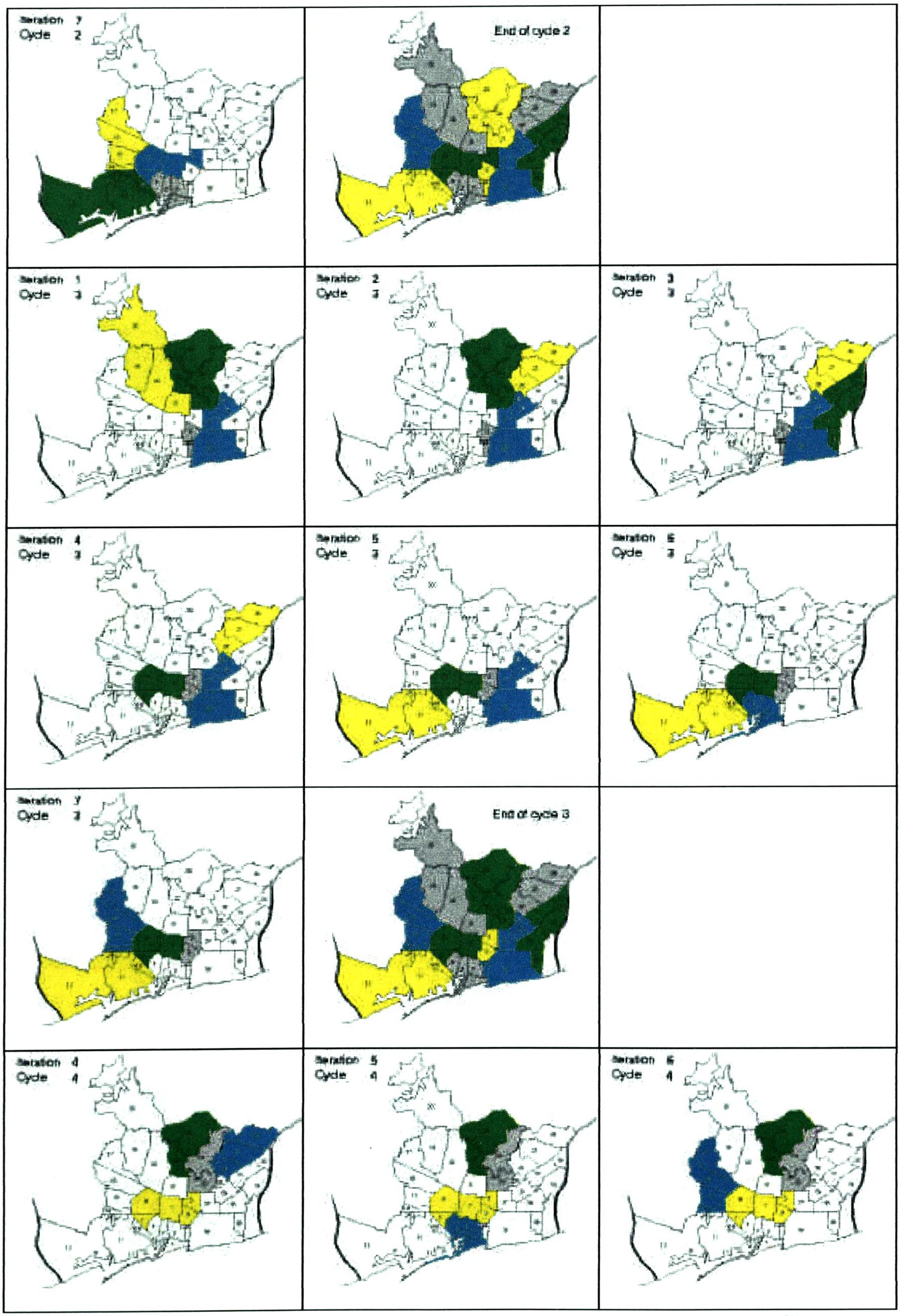
Tesis Doctoral para optar al título de Doctor en Estudios Empresariales
Director: Dr. Manuel Artís Ortuño
Doctorado en Estudios Empresariales
Programa: Técnicas y Análisis en la Empresa
Bienio 2000-2002

4.6. ANNEX

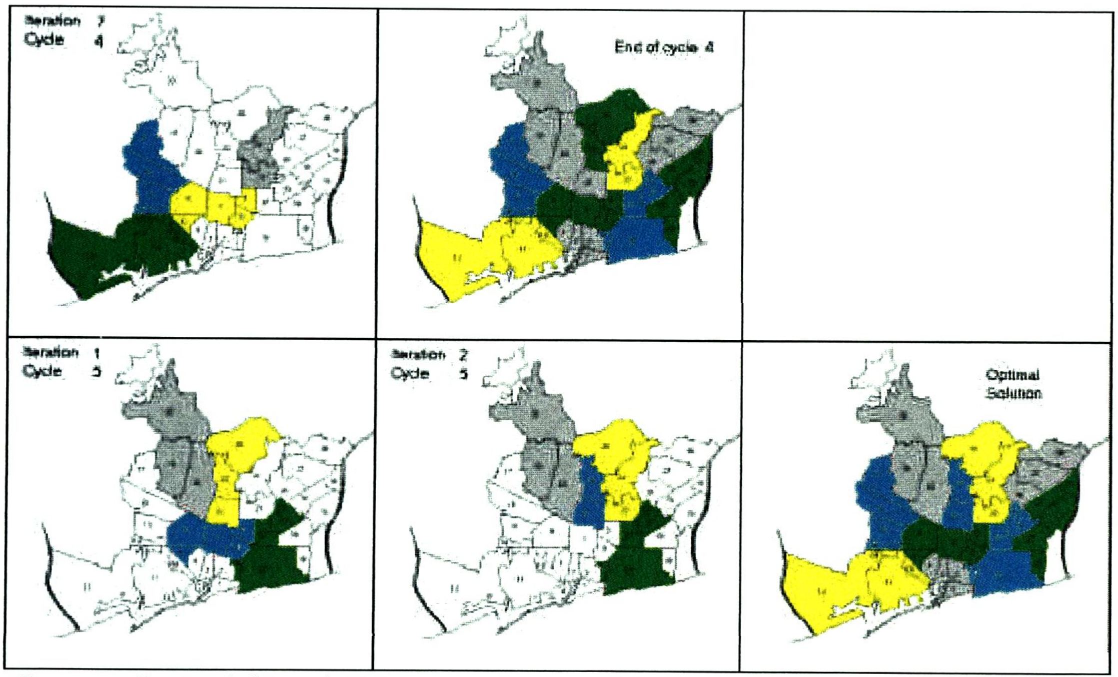
Maps of the different territorial configurations obtained using RASS.



Continue...


Iteration 7
Cycle 2
End of cycle 2
Iteration 1
Cycle 3
Iteration 2
Cycle 3
Iteration 3
Cycle 3
Iteration 4
Cycle 3
Iteration 5
Cycle 3
Iteration 6
Cycle 3
Iteration 7
Cycle 3
End of cycle 3
Iteration 4
Cycle 4
Iteration 5
Cycle 4
Iteration 6
Cycle 4

Continue...



Source: Own elaboration.

CHAPTER 5

An empirical illustration of the proposed methodology in the context of regional unemployment in Spain

5.1. Introduction.

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 to the phenomena analysed.

There are many economic variables whose analysis at a nationwide aggregation level is not representative because of large-scale regional disparities. These regional disparities make it necessary to complement the aggregated analysis with applied research at a lower aggregation level in order to have a better understanding of the phenomenon being studied. A clear example of this case can be found when analysing the Spanish unemployment rate. Previous studies have demonstrated that Spanish unemployment presents major disparities (Alonso and Izquierdo, 1999), accompanied by spatial dependence (López-Bazo *et al.*, 2002) at the provincial aggregation level (NUTS I). In fact, these two elements, disparity and spatial dependence, make this variable a good candidate for regionalisation experiments that allow the study of the differences that can be generated between the normative and analytical geographical divisions. The analysis in this chapter focuses on quarterly provincial unemployment rates in peninsular Spain from the third quarter of 1976 to the third quarter of 2003. Table 5.1 shows the NUTS classification for the Spanish regions.

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

Table 5.1. NUTS Classification for the Spanish regions.

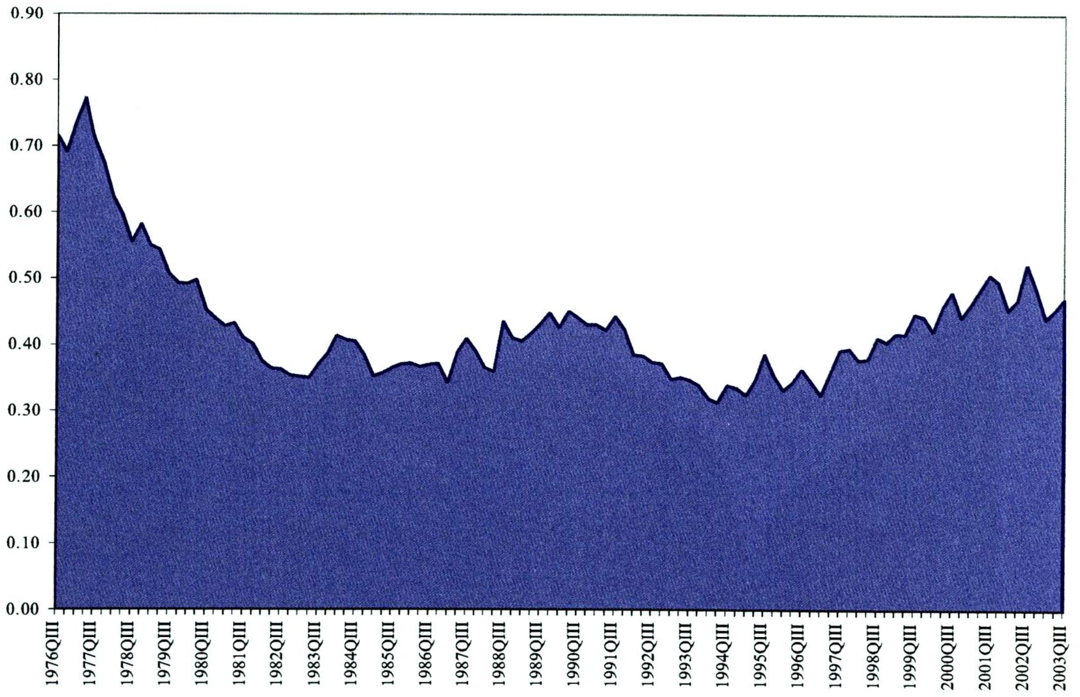
NUTS I	NUTS II	NUTS III	CODE
NOROESTE	GALICIA	Coruña (A)	16
		Lugo	27
		Orense	32
		Pontevedra	34
	ASTURIA	Asturias	5
	CANTABRIA	Cantabria	12
NORESTE	PAIS VASCO	Álava	1
		Guipúzcoa	21
		Vizcaya	45
	NAVARRA	Navarra	31
	RIOJA	Rioja (La)	35
	ARAGON	Huesca	23
		Teruel	41
		Zaragoza	47
MADRID	MADRID	Madrid	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

5.2. Regional Unemployment in Spain: spatial differences and dependence.

As regards spatial disparity, Figure 5.1 shows the variation coefficient of NUTS III unemployment rates during the period considered. As can be seen, throughout the period, there is a major 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 5.1. Variation coefficient for the unemployment rate at NUTS III level.



Source: Own elaboration

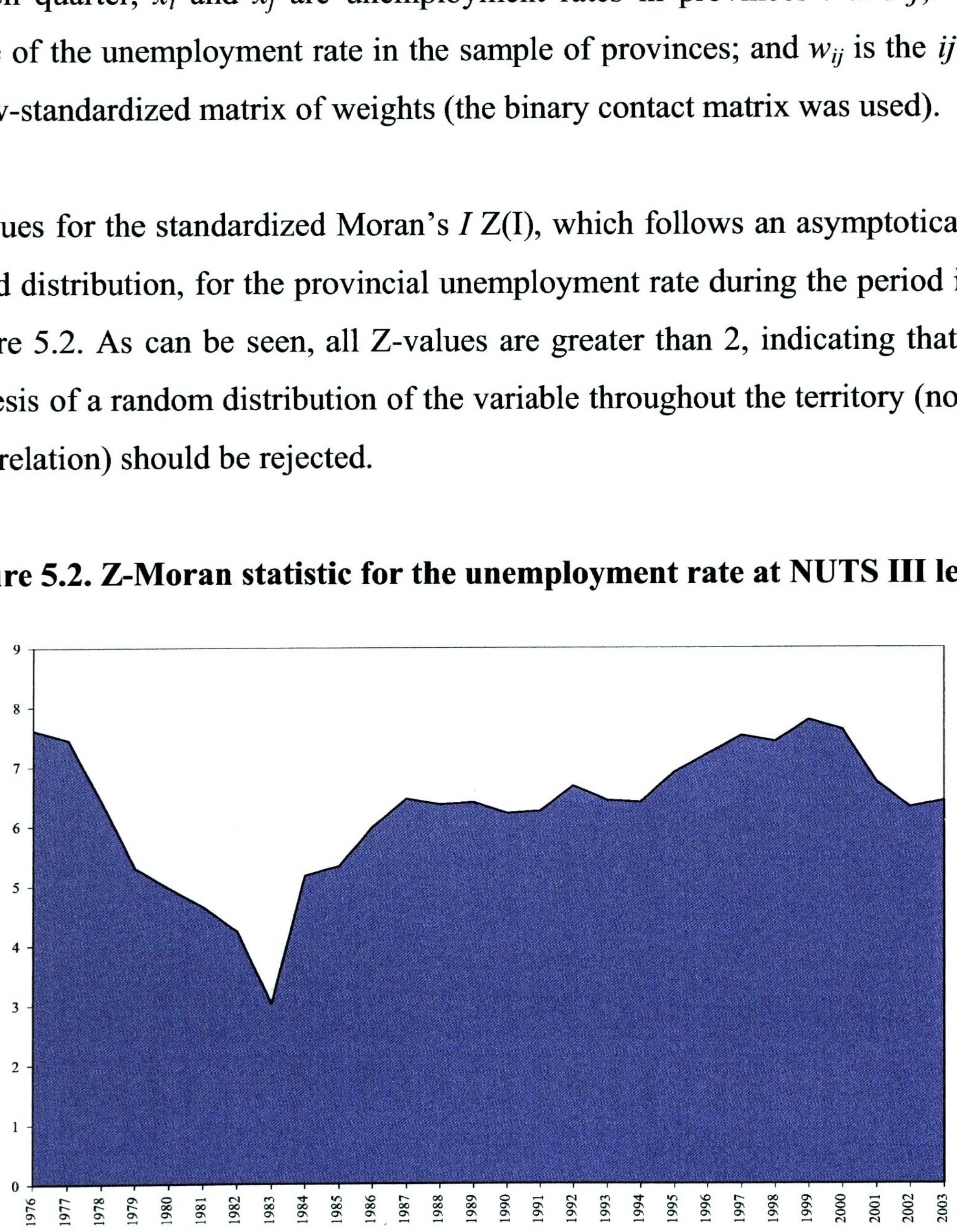
For spatial dependence, the Moran's *I* statistic (Moran, 1948) of first-order spatial autocorrelation has been calculated.

$$I = \frac{\sum_{ij}^{N} w_{ij}\left(x_i - \bar{x}\right)\cdot\left(x_j - \bar{x}\right)}{\left(x_i - \bar{x}\right)^2} \qquad i \neq j \qquad (5\cdot 1)$$

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-standardized matrix of weights (the binary contact matrix was used).

The values for the standardized Moran's I Z(I), which follows an asymptotical normal standard distribution, for the provincial unemployment rate during the period is shown in Figure 5.2. As 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 5.2. Z-Moran statistic for the unemployment rate at NUTS III level[31].



Source: Own elaboration

[31] 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.

This descriptive analysis shows that a regionalisation process is clearly justified: The existence of spatial differences gives rise to the creation of groups, whereas the spatial dependence supports the imposition of geographical contiguity of these groups.

5.3. Normative regions: NUTS classification.

To compare the results obtained using analytical regionalisation procedures or using the territorial division NUTS, which were established according to normative criteria, we will now design the regions on the basis of the behaviour of provincial unemployment, ensuring that provinces belonging to the same region are as homogeneous as possible in terms of this variable.

To aid the comparison with the NUTS division, we establish two scale levels. The first comprises 15 regions for comparison with peninsular Spain's 15 NUTS II regions, and the second has six, for comparison with peninsular Spain's 6 NUTS I regions.

One way of comparing the homogeneity[32] of the different territorial divisions is to calculate Theil's inequality index (Theil, 1967). One advantage of this index in this context is that its value can be broken down into a within-group component and a between-group component.

$$T = \sum_{p=1}^{n} \frac{u_p}{U} \log \left[\frac{\left(\frac{u_p}{U} \right)}{\left(\frac{1}{n} \right)} \right] \qquad (5\cdot 2)$$

[32] 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.

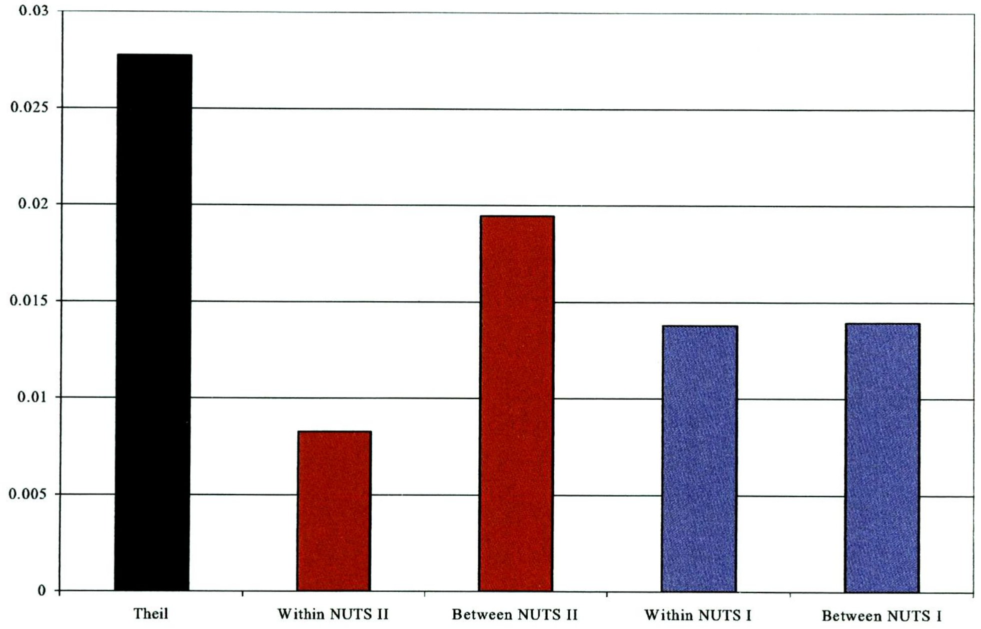
Where n is the number of provinces (47), u_p is the provincial unemployment rate indexed by p, and U represents the Spanish unemployment rate $U = \sum_{p=1}^{n} u_p$

Overall inequality can be completely and perfectly broken down 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 only region to which the province belongs, and subscript p, where, in each region, p goes from 1 to n_i.

The aim of analytical regionalisation procedures is to minimise within inequalities and maximise between inequalities.

Figure 5.3 shows the total value of the Theil inequality index and the value of the within-group and between-group components when average unemployment rates of Spanish provinces (NUTS III) are aggregated into NUTS II and NUTS I regions. The most important result from this figure is that the level of "internal" inequality (the within component) is very high (in relative terms) for both scale levels, but in particular for the NUTS I level.

Figure 5.3. Decomposition of Theil's index for the average unemployment rate (from 1976-QIII to 2003-QIII) for NUTS III into NUTS II and NUTS I regions.

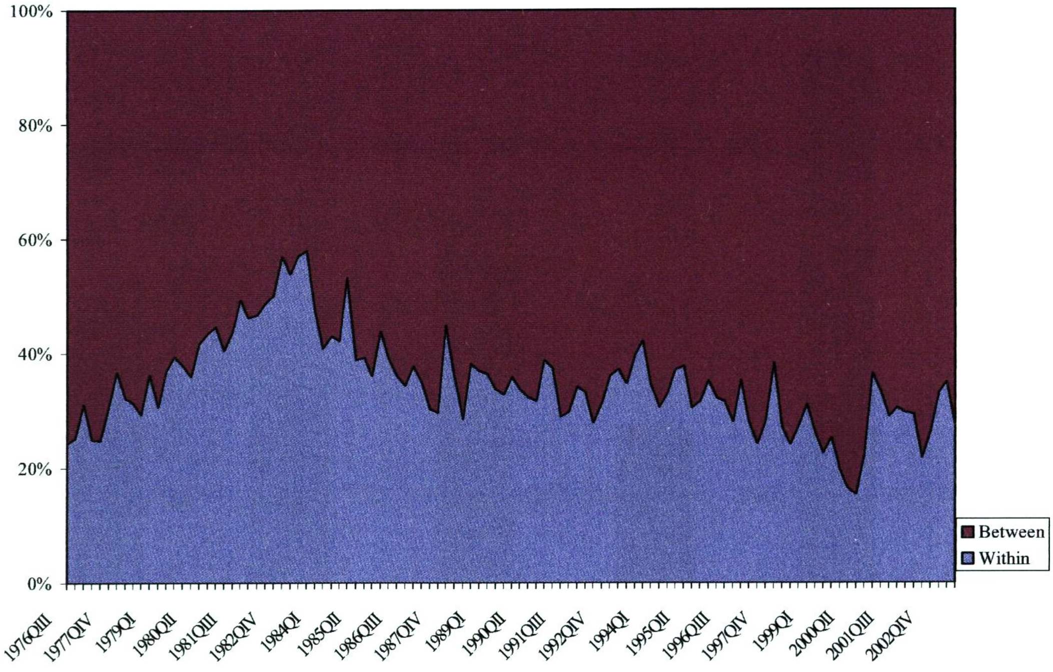


Source: Own elaboration

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, as far as the provincial unemployment rate is concerned, are the NUTS regions representative of the behaviour of regional unemployment during the whole period? Figures 5.4 and 5.5 show the relative decomposition of Theil's inequality index throughout the period analysed. For both, NUTS II (Figure 5.4) and NUTS I (Figure 5.5) it can be seen that the behaviour of the "within" inequality is irregular, with its greatest dispersion at the beginning of the eighties. The highest homogeneity level is reached during 2000. Note also that the proportion of "within" inequality in NUTS I is much higher than 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 NUTS II to obtain NUTS I[33].

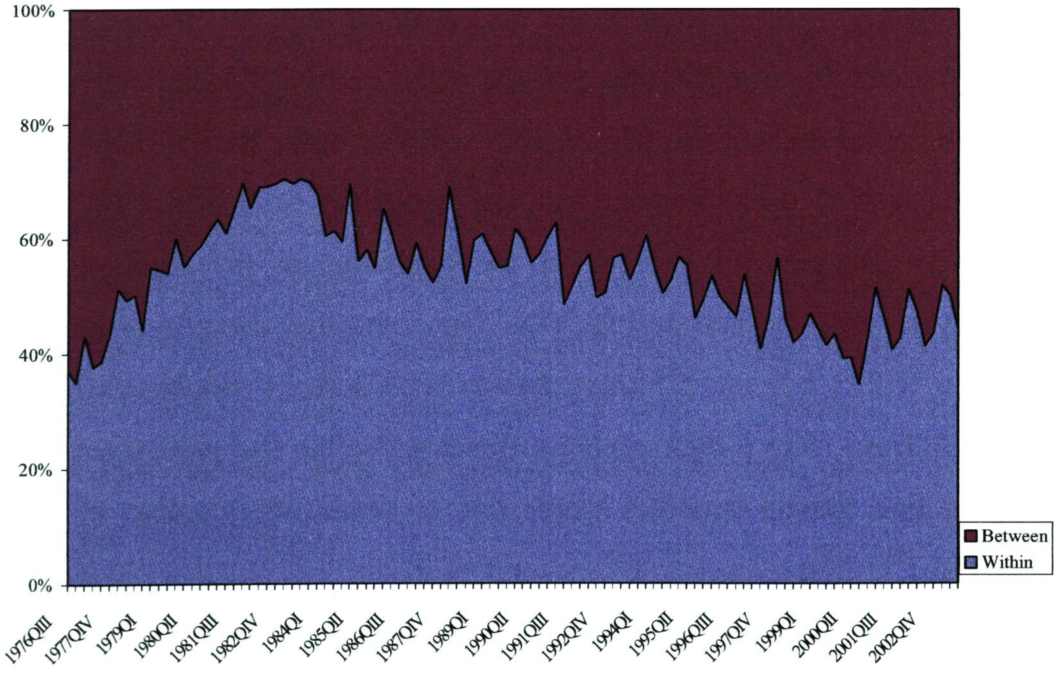
[33] This disadvantage was discussed above, in chapter 2, when hierarchical aggregation was introduced.

Figure 5.4. Decomposition of Theil's index for the unemployment rate for NUTS III regions into NUTS II regions.



Source: Own elaboration

Figure 5.5. Decomposition of Theil's index for the unemployment rate for NUTS III regions into NUTS I regions.



Source: Own elaboration

5.4. Normative vs. analytical regions.

Can an analytical regionalisation process improve the results obtained for normative regions? To answer this question, we applied a "two-stage" regionalisation strategy based on K-means algorithm and the *RASS* algorithm.

The K-means algorithm is applied to the unemployment rates to group the 47 contiguous provinces into 15 and 6 regions. The results will be compared with the normative regions (NUTS II and NUTS I) presented above. The same process will also be performed by applying the *RASS* algorithm. Finally, we compare the K-means and the *RASS*.

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 the regionalisation process with a dynamic component for the temporal design of representative regions. The use of euclidean distances (squared in K-means) allows us to take into account the differences in both direction and magnitude between the unemployment rates in the different areas.

Figure 5.6 shows a comparison between normative and analytical regions using K-means. The values below the provincial code indicate the deviation from the (unweighted) arithmetical average of the unemployment rate of the region to which it belongs[34]. It is expected that if regions are homogeneous, then the provincial unemployment rate will be near the regional one.

For NUTS II (upper map) the maximum deviations are located in Barcelona (number 8 in the map) with a rate 6.06% above the regional average, and Almería (4), with a rate

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

7.83% below the regional average. Note that the range is 13.88, which indicates substantial differences in the unemployment rate between provinces inside the same region.

With respect to analytical regions obtained by K-means (bottom 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 considerably lower than before.

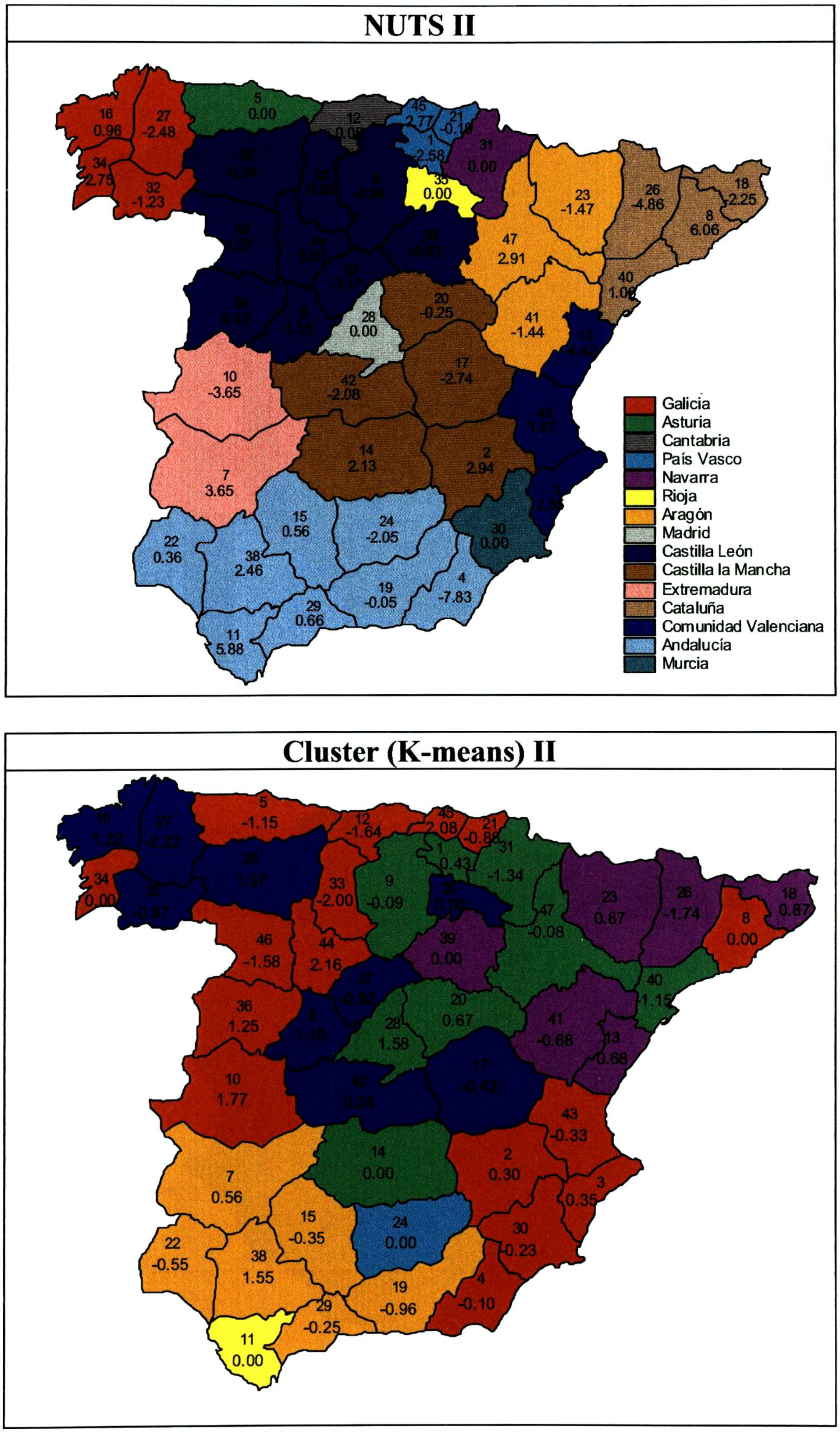
After designing 15 analytical aggregations for comparison with NUTS II, the unemployment rate is re-calculated for each of the 15 regions. The new series are used to aggregate those 15 regions into 6 analytical regions. This method ensures that the aggregation obtained is nested inside the previous one so as to allow comparison with NUTS I. It is important to note that when the K-means cluster is applied, it is impossible to obtain six regions, because when the number of cluster regions is set at three, the number of contiguous regions is seven[35].

Figure 5.7 shows the normative regions (upper map) that correspond to NUTS I aggregation level, and analytical regions (bottom 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% (Navarre - 31). The range has now decreased from 17.93 to 8.25.

Table 5.2 shows the results of the regionalisation process using the K-means cluster procedure.

[35] If the value of the cluster regions is set at two, then only two contiguous regions are obtained.

Figure 5.6. Comparison between administrative (NUTS II) and economic regions using the K-means cluster.

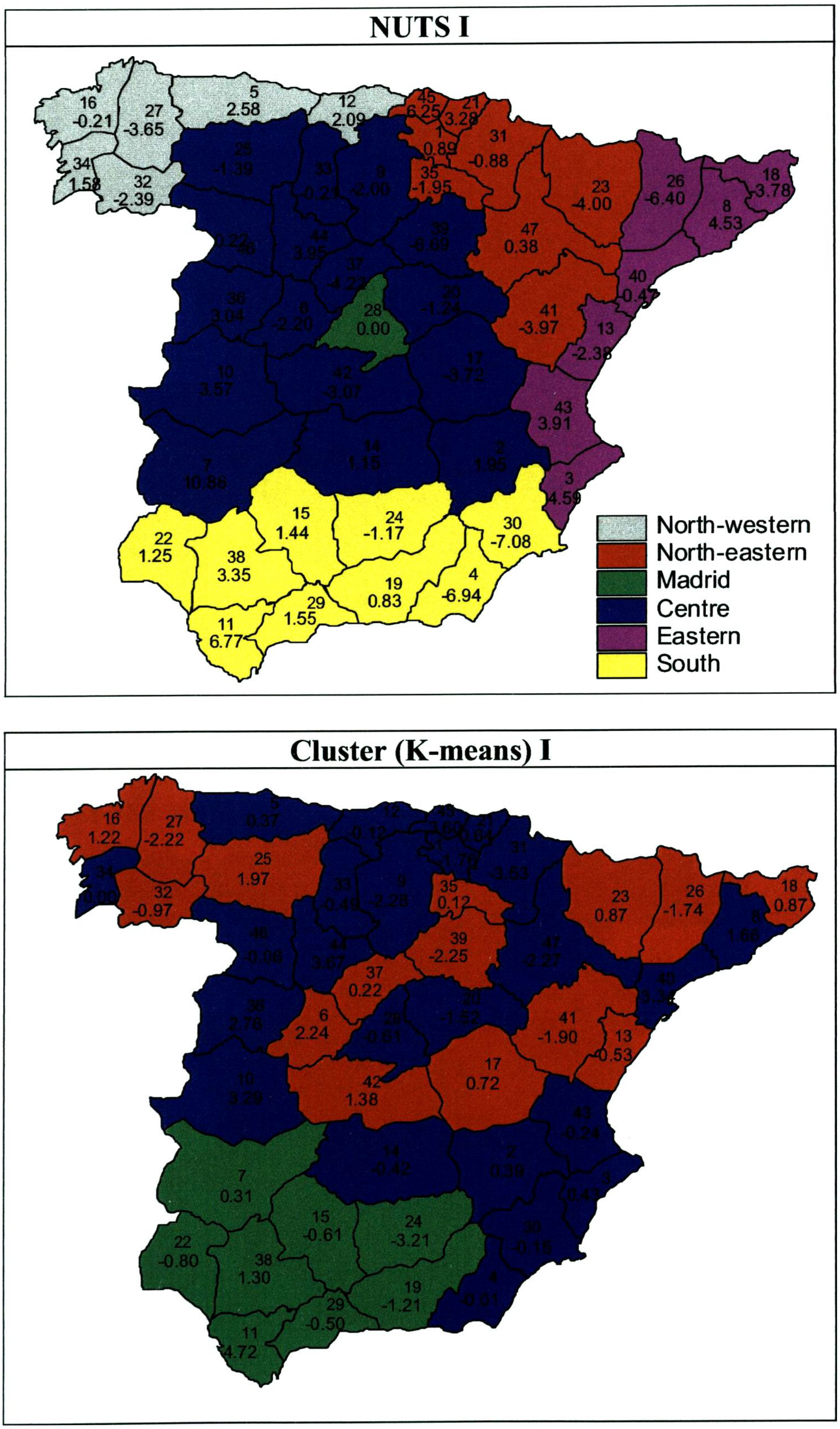


NUTS II

Cluster (K-means) II

Source: Own elaboration

Figure 5.7. Comparison between administrative (NUTS I) and economic regions using the K-means cluster.



NUTS I

Cluster (K-means) I

Source: Own elaboration

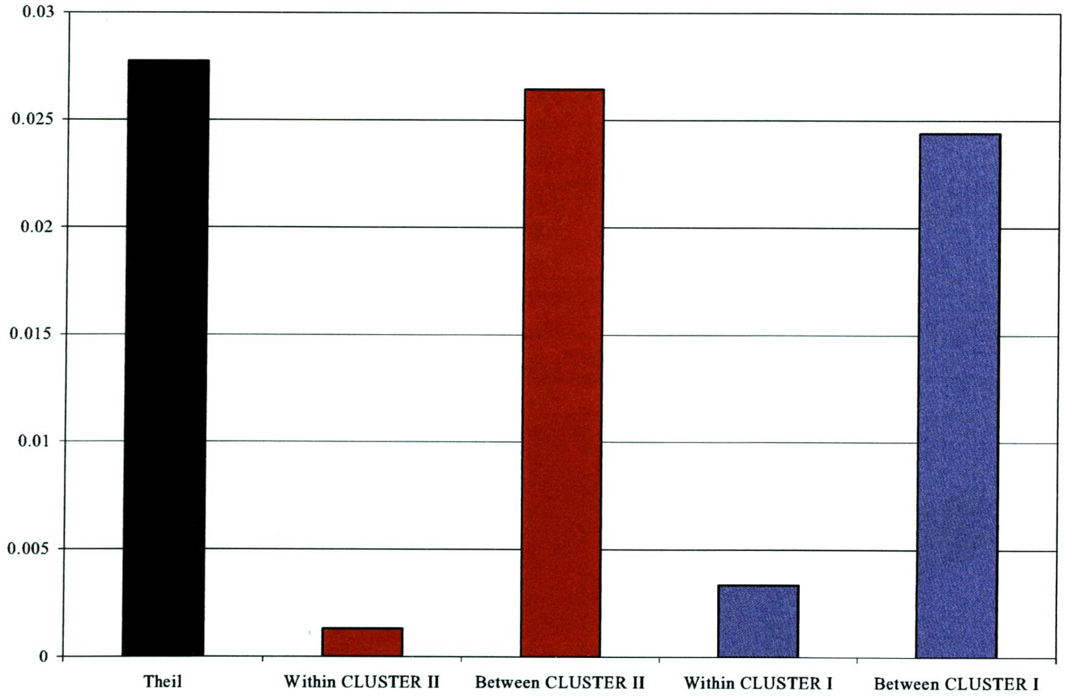
Table 5.2. Detailed results of the regionalisation process using the K-means cluster procedure.

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

For a more detailed analysis of the homogeneity reached by using analytical regionalisation with the K-means algorithm, Theil's inequality index was again calculated. The results in Figure 5.8 show a major improvement in terms of within/between inequality. In both cases, CLUSTER II and CLUSTER I aggregation levels, inequality within regions represents only 4.68% and 11.98% of the total inequality between provinces. This implies that analytical regions are much more homogeneous than normative ones in terms of average unemployment rates.

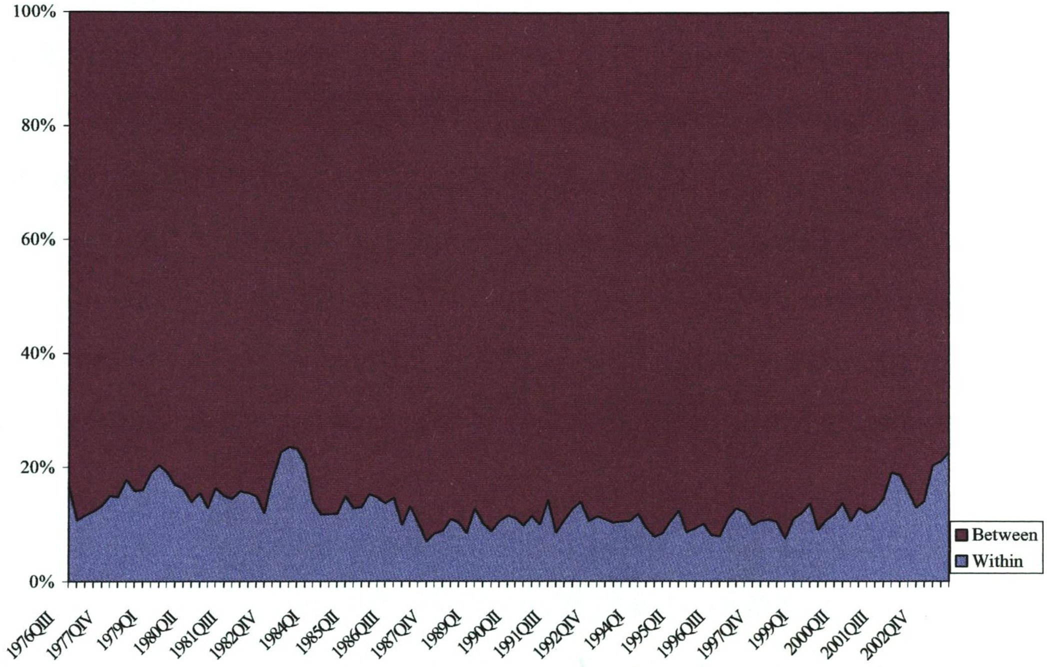
Figure 5.8. Decomposition of Theil's index for the unemployment rate for NUTS III regions into Cluster II and Cluster I regions.



Source: Own elaboration

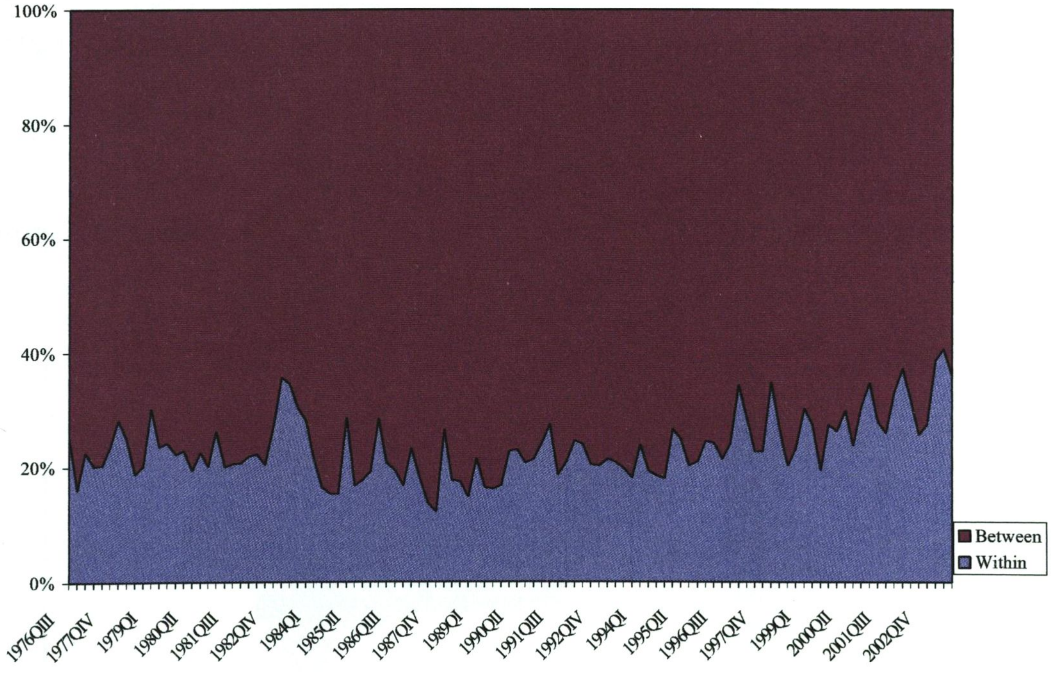
Another important result is obtained when Theil's inequality index is calculated for each quarter for the different aggregation levels (Figures 5.9 and 5.10). As can be seen, the "within" inequality is more constant for analytical regions than for normative regions.

Figure 5.9. Decomposition of Theil's index for the unemployment rate for NUTS III regions into Cluster II regions.



Source: Own elaboration

Figure 5.10. Decomposition of Theil's index for the unemployment rate for NUTS III regions into Cluster I regions.



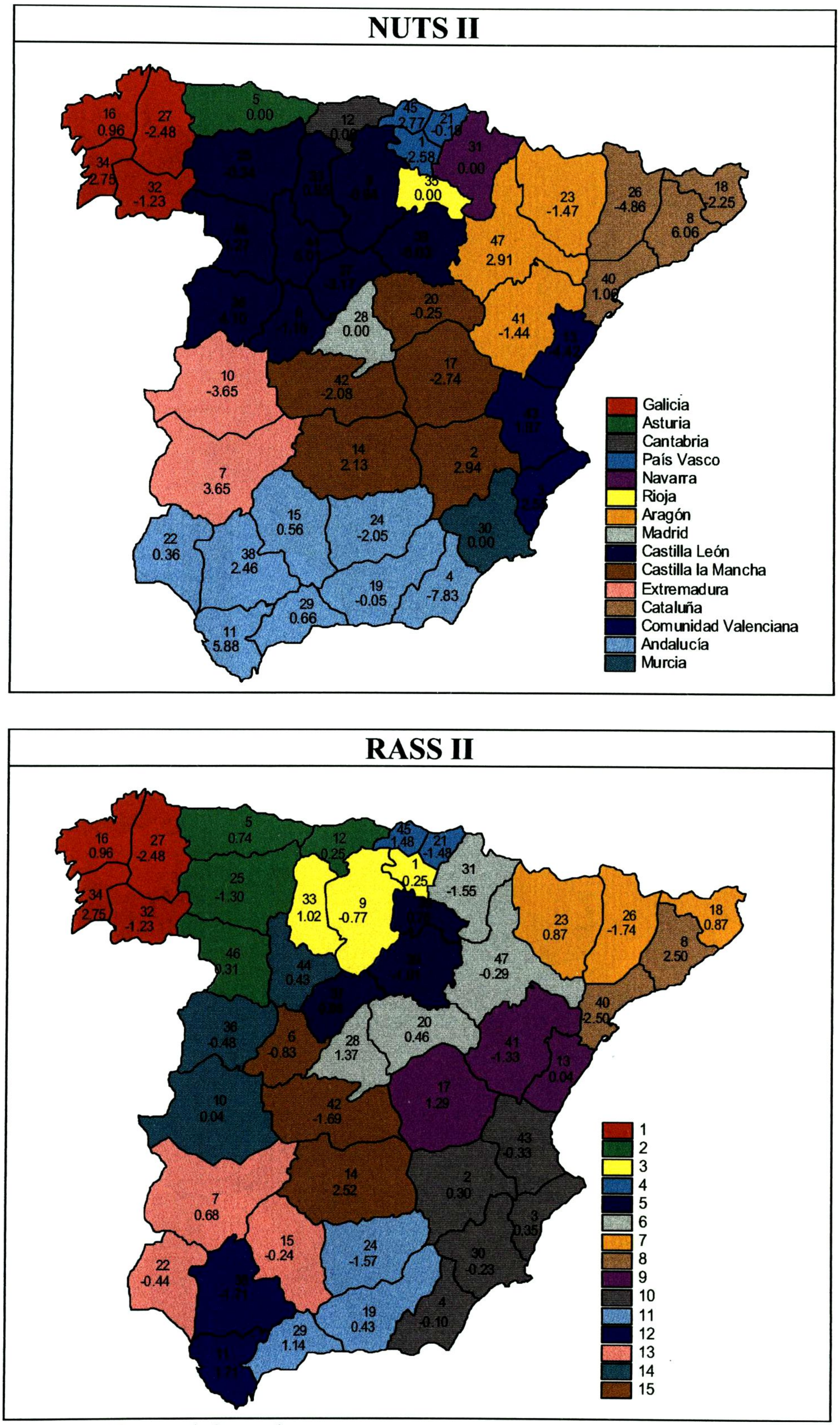
Source: Own elaboration

The second analytical regionalisation procedure applied in this paper is the *RASS* algorithm. Figures 5.11 and 5.12 show the analytical regions obtained by applying *RASS* and the normative regions (NUTS) for the two aggregation levels considered. At both levels, the average unemployment rates show lower deviations with respect to regional averages when using *RASS*. In RASS II, Pontevedra (34) present the highest deviations (2.75%) and Tarragona (40) the lowest (-2.50%). In the RASS I aggregation, the extreme deviations are located in Barcelona (8) and Lleida (26) with deviations from regional averages of 6.51% and -4.42% respectively. In both cases, the ranges are considerably lower in the *RASS* regions than in normative regions, as in the K-means case.

The values of Theil's inequality index (Figure 5.13) calculated for RASS II and RASS I regions using the average unemployment rates show that the inequality within regions is strongly reduced to 6.54% and 21.64% of the total inequality. This suggests again that analytical regions using the *RASS* are much more homogeneous than normative ones in terms of average unemployment rates. In RASS II, the "within" inequality remains relatively constant over the period analysed (Figure 5.14), but for RASS I (Figure 5.15) the "within" inequality is especially high between 1976 and 1984.

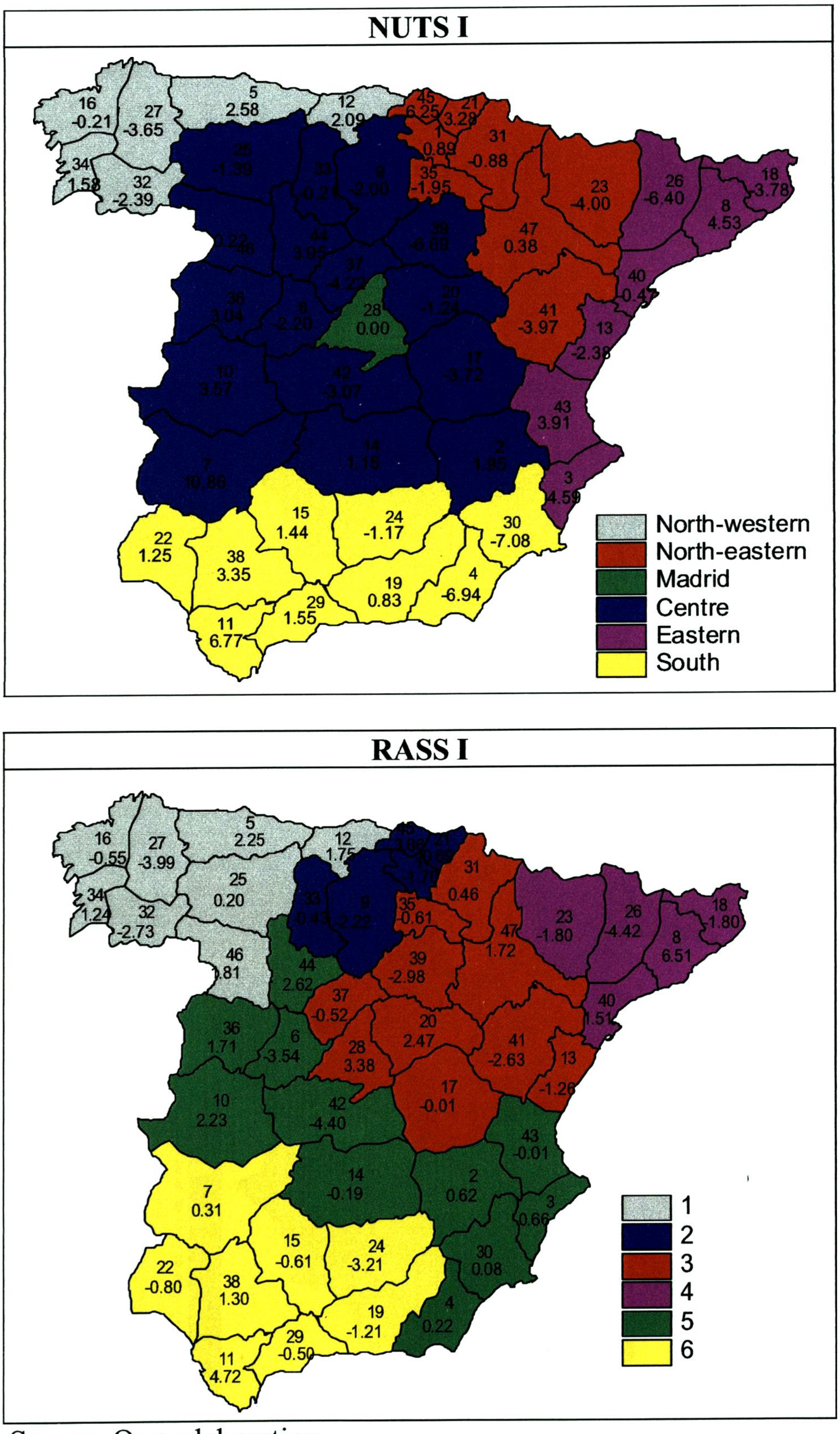
Detailed results of the regionalisation process using the *RASS* procedure are shown in Table 5.3.

Figure 5.11. Comparison between administrative (NUTS II) and economic regions using the RASS procedure.



Source: Own elaboration

Figure 5.12. Comparison between administrative (NUTS I) and economic regions using the RASS procedure.



NUTS I

RASS I

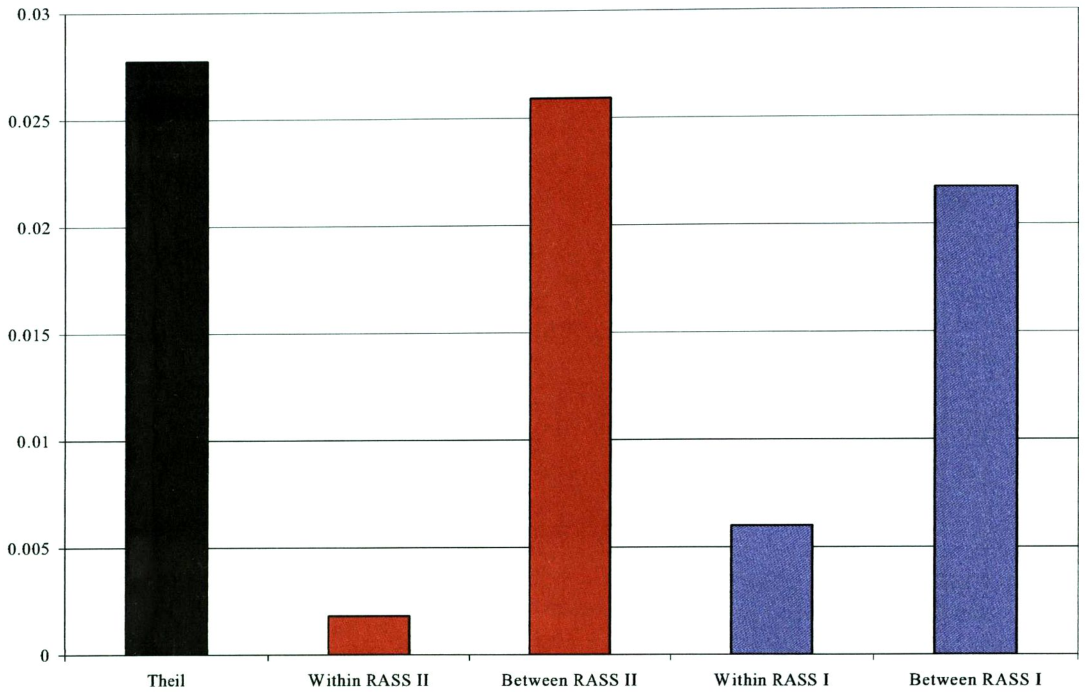
Source: Own elaboration

Table 5.3. 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

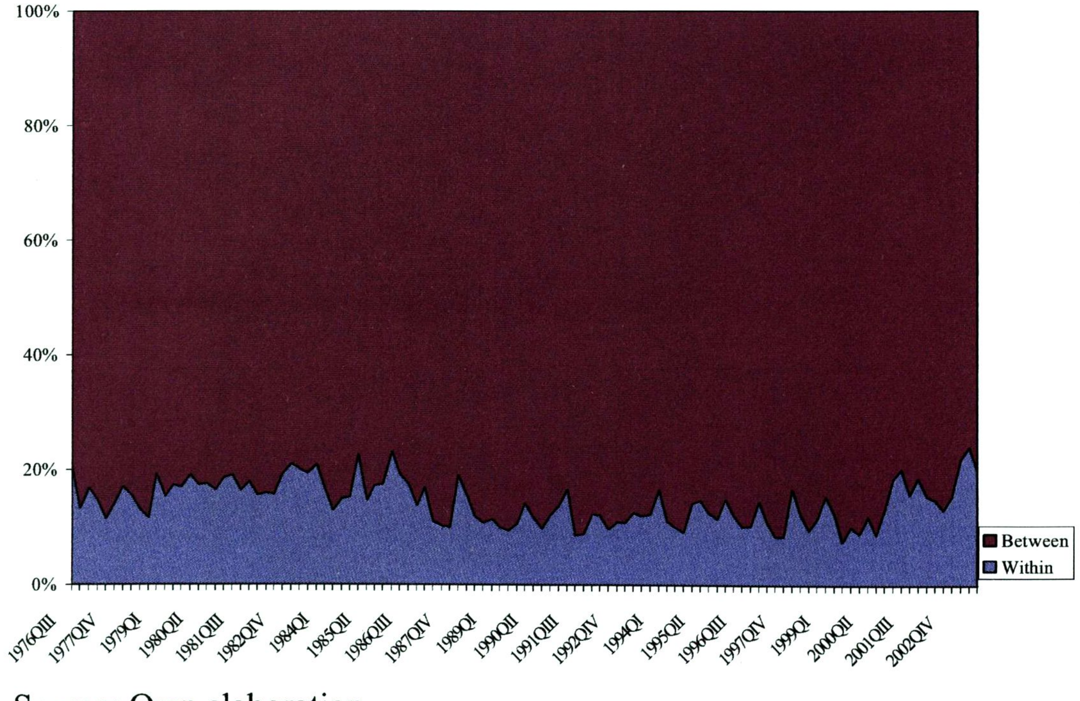
Source: Own elaboration

Figure 5.13. Decomposition of Theil's index for the unemployment rate for NUTS III regions into RASS II and RASS I regions.



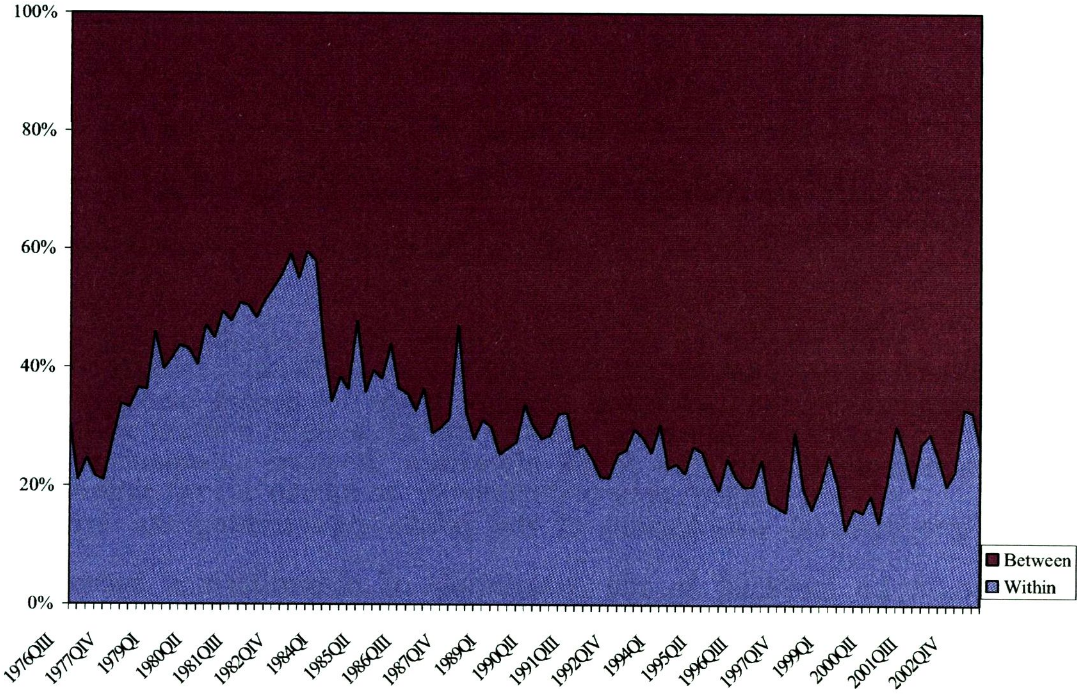
Source: Own elaboration

Figure 5.14. Decomposition of Theil's index for the unemployment rate for NUTS III regions into RASS II regions.



Source: Own elaboration

Figure 5.15. Decomposition of Theil's index for the unemployment rate for NUTS III regions into RASS I regions.



Source: Own elaboration

5.5. Final remarks.

Table 5.4 summarises the basic descriptive statistics discussed above. These statistics establish the basis for a comparison between the different regionalisation procedures applied. The comparison is divided into different regionalisation characteristics: Homogeneity, regional shape, control level and flexibility. In each category the main advantages or disadvantages of each analytical method are mentioned.

Homogeneity: Both analytical regionalisation methods greatly improve the intra-regional homogeneity throughout the period. For both aggregation levels (II and I), Clustering method (using K-means algorithm) obtains lower values of within-region dispersion (see Table 5.3). Note that in the CLUSTER II aggregation level seven regions are composed by only one area; this could account for the lower value of "within" inequality.

Table 5.4. 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, the two-stage strategy tends to design strongly irregular region shapes compared with the *RASS* algorithm. 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[36]. Also, with the two-stage strategy, the number of provinces grouped in each region shows large differences: in CLUSTER II there are seven regions formed by one province, but there are also regions formed by nine provinces. The same happens in CLUSTER I, since the number of provinces assigned to a region ranges between one and seventeen. On the other hand, *RASS* algorithm forms more balanced regions: at RASS II, the number of provinces by regions varies between two and four, and between five and eleven at RASS I.

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

[36] For a more detailed discussion of this problem, see Wise *et al.*, 1997.

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

Finally, researchers cannot ignore the fact that the results are sensitive to the type of regionalisation method applied. In this context, analytical regionalisation models are a good alternative for the design of representative geographical units.