



XLème Colloque de l'ASRDLF

## Convergence et disparités régionales au sein de l'espace européen

### Les politiques régionales à l'épreuve des faits

Bruxelles – 1, 2 et 3 Septembre 2004

## Connectropy and Cluster Analysis: Results Compared.

Jean H.P. Paelinck  
George Mason University  
School of Public Policy\*

### *Summary.*

In previous papers the two types of analysis mentioned - connectropy and cluster analysis - have been applied to Belgian and Dutch regional units to generate macro-regions. This paper compares the results and tries to discover reasons for matching or diverging results.

*JEL classification:* C4, C5, R1.

*Table of contents.*

1. Introductions.

2. Methodology.

2.1. Connectropy.

2.2. Clustering.

3. Applications and comparison.

3.1. The Netherlands.

3.2. Belgium.

4. Conclusions.

5. References.

6 Appendix: maps.

## 1. Introduction.

In previous papers (Kaashoek, Paelinck and Zoller, 2004; Paelinck, 2004) we developed two different methods to construct regional groupings. One was based on interregional linkages, using the concept of connectropy; the other one presented an algorithm to cluster elementary regional units in a contiguous way. Both methods were applied to Dutch and Belgian regions; section 3 compares those results, after section 2 has briefly summarised the two methods used. Conclusions and references then follow.

## 2. Methodology.

### 2.1. Connectropy.

One starts with an axiomatic definition of "nearness" or "proximity", along the lines of that of a distance measure or metric.

Take any set of elements,  $S$ , and consider three of those elements,  $x$ ,  $y$  and  $z$ . Define on two of them a numerical function,  $n(\dots)$ , which obeys the following properties:

$$n(x,y) = 1, x=y \quad (1)$$

$$0 < n(x,y) < 1, x \neq y \quad (2)$$

$$n(x,y) = n(y,x) \quad (3)$$

$$n(x,y) \bullet n(y,z) \leq n(x,z) \quad (4)$$

Function  $n(\dots)$  will then be called a "nearness" measure; the similarity with an axiomatic definition of a metric is obvious, it being stressed that the well-known triangular inequality is now stated in terms of a product and a  $\leq$ -relation.

A possible and useful member of the family of nearness (or closeness) measures is the following:

$$n(i,j) = \exp(-d_{ij}) \quad (5)$$

where  $d_{ij}$  is any proper distance measure between entities  $i$  and  $j$ , e.g. the degree of contiguity, which is indeed a proper distance measure, and which will be used later on; one can easily check that (5) is a proper closeness measure, given the exponential additivity property and the triangular inequality property of metrics.

Consider now the function:

$$c_{ij} = d_{ij} \exp(-d_{ij}) \equiv -n(i,j) \bullet \ln n(i,j) \quad (6)$$

of which the similarity with an entropy measure is again obvious: indeed read "probability" for variable  $n$ , and the formal likeness is complete.

But also from the intrinsic point of view this is true; equation (6) can indeed be given an informational content. The occurrence of a long distance relation is relatively rare, so its observation contains a high degree of information, which is measured here by  $d_{ij}$ ; on the other hand, it is known from distance interaction that the effective intensity of interaction declines

with distance [in (6) in a negative exponential way, "radioactive decay"], and as entropy measures the expected information, so (6), which we will call "connectropy", measures the expected intensity.

Results of such a measuring can be treated as follows: suppose one avails of a matrix of connectropy measures; one could then try to extract groups with certain values of total connectropy,  $\sum_{i,j} c(i,j)$  (to be compared to total entropy), e.g. declining levels from a maximum to a minimum; this amounts in fact to applying linear assignment to the matrix, as will be shown now.

Let  $\mathbf{A}$ , of order  $n \times n$ , be that matrix; note that if the diagonal is not void, its terms have been put equal to zero in all the exercises performed.

One could try and generate several paths, possibly generating cyclical groups (or circuits), in the sense that the last entity connects with the first one, each maximising the sum total of the interrelationships present, the objective being to isolate *several* groups with declining maximal total internal intensity, and study the resulting interconnections.

The *initial* mathematical program can then be written as:

$$\max_{\mathbf{x}} \varphi = [\mathbf{vec}(\mathbf{A}')]'\mathbf{x} \quad (7)$$

subject to:

$$\mathbf{J}\mathbf{x} \leq \mathbf{i} \quad (8)$$

$$\hat{\mathbf{x}}\mathbf{x} = \mathbf{x} \quad (9)$$

$\mathbf{i}$  being the unit column vector.

The following definitions and interpretations should be given:

\*  $\mathbf{vec}(\mathbf{A}')$  is the vectorisation of the transpose of  $\mathbf{A}$ , all diagonal elements omitted, so of order  $n(n-1) \times 1$ ;

\*  $\mathbf{x}$  is a vector of order  $n(n-1) \times 1$  of binary variables [conditions (9), where  $\hat{\mathbf{x}}$  is the diagonal matrix constructed from  $\mathbf{x}$ ];

\* conditions (8) are weakened assignment conditions, matrix  $\mathbf{J}$  being binary and of order  $2n \times n(n-1)$ ; it is the so-called assignment matrix, assigning a unique place to each of the units taken up in the solution; if the weak inequalities were to be replaced by equalities, exactly  $n$  *directed* relations would be selected, each agent appearing twice, so relaxation allows of generating *incomplete groups*.

In fact, the mathematical program (7) through (9) is the inverse (a maximum instead of a minimum) of the traveling salesman problem, but here the traveling salesman is in no hurry at all, and not obliged to pass through all the potential nodes, and moreover he is allowed to go back and forth and to travel along disconnected paths!

Once the solution to the mathematical program has been obtained (by simple linear programming) one has generated a first group with maximal internal cohesion, meaning here

total connectropy; one then cancels the corresponding entries of  $\mathbf{A}$ , leading up to a matrix  $\mathbf{A}^*$  to be treated in the same manner, and so on until all the entries have been exhausted..

Symmetry of matrix  $\mathbf{A}$  leads to symmetrical flows in the optimal solution, which can be prevented by introducing the conditions:

$$x_{ij} + x_{ji} \leq 1, \forall i < j \quad (10)$$

but they would still not guarantee complete connectivity; such a requirement would in fact lead to a quadratic assignment problem..

## 2.2. Clustering.

The model here runs as follows.

Specify an objective function, to be minimised, as :

$$\varphi = \sum_{i,j>i} d_{ij} (x_{i1}x_{j1} + \dots + x_{ik}x_{jk}) \quad (11)$$

where  $i$  and  $j$  are indices for  $n$  observations, and  $d_{in}$  entries of a distance matrix,  $\mathbf{D}$ , derived from the values of the observations; the  $x_{ic}$ ,  $x_{jc}$ ,  $1 \leq c \leq k$ , are binary decision variables assigning observations to  $k$  regimes. A value  $d_{ij}$  is only effective if both observations  $i$  and  $j$  belong to the same regime.

If  $\mathbf{X}$  is then the  $k \times n$  matrix constructed from the row vectors of the  $x_{ic}$ s, a first constraint is:

$$\mathbf{X}\mathbf{i} = \mathbf{i} \quad (12)$$

the  $\mathbf{i}$ 's being conformable unit column vectors; the meaning is that each  $x_{ic}$  should be allocated to one single regime. Next constraint is:

$$\mathbf{X}'\mathbf{i} \geq \mathbf{i} \quad (13)$$

meaning that each cluster should contain at least one observation. One extra constraint could be:

$$x_{ic} = 1 \quad (14)$$

meaning that observation  $i$  belongs to cluster  $c$ , where  $i$  and  $c$  are arbitrary.

Each product in (11) can be linearised as follows; take for instance the product  $x_{ic}x_{jc}$ , and write:

$$d_{ij}(x_{ic} + x_{jc} - 1) + y_{ijc} \geq 0 \quad (15)$$

where  $y_{ijc}$  is a real auxiliary variable; all possible values of the product  $x_{ic}x_{jc}$  can be matched by (15). Indeed, if  $x_{ic}$  and  $x_{jc}$  are both equal to one,, the value of (15) will be one, as  $y_{ijc}$ , appearing in the objective function to be minimised, will be zero. If either  $x_{ic}$  or  $x_{jc}$  is zero, the value of (15) will be zero,  $y_{ijc}$  being again zero. Finally for both  $x_{ic}$  and  $x_{jc}$  equal to zero, (15) will again be zero with this time  $y_{ijc} = d_{ij}$ , the smallest value it can take for (15) to be non-negative.

The disadvantage of the specification just presented is that it introduces extra  $y_{ijc}$  variables (but those are also present in other specifications); this specification is however useful as it can be solved by linear programming (specifying  $x_{ic} \geq 0, \forall i,c$ ) and can be generalised to Quadratic Assignment Problems .

In time series, continuity restrictions have to be imposed; in spatial analysis, contiguity restrictions are to be introduced in order to generate "continuous" regions. Those restrictions are:

$$\mathbf{x}_c' \mathbf{C}_1 \mathbf{x}_c \geq \mathbf{i}' \mathbf{x}_c - 1 \quad (16)$$

Indeed, the left hand term is the sum of the relevant cross-products for cluster  $c$ , and that number should at least be equal to the total number of terms minus one.

It has - as yet - not been possible to find a linear analogon to (16), so a differential solution procedure was devised.

For problems of the type investigated, a simple solution procedure can indeed be devised. Indeed, the distances appearing in the distance matrix  $\mathbf{D}$  are "natural" distances, i.e. distances along a line. So, classifying the items to be investigated from smallest to highest values, non-decreasing series are generated; the relative increases are present in terms of rows and columns of  $\mathbf{D}$ . Inspection of the column sums of the values above the main diagonal of  $\mathbf{D}$  will show jumps at certain points; from there on, a new cluster should be investigated.

As to contiguity condition (16), it can be treated the same way, starting from the smallest admissible (contiguous) values; the procedure is then the same for the permuted matrix.

An index,  $I_c$ , can be computed as  $\varphi/\varphi_c$ , where  $\varphi$  is the value of the free optimum,  $\varphi_c$  being the contiguity-constrained one; it is an indicator of the "strength" of the contiguity condition, a value of 1 meaning that the constraint is superfluous, a value  $I_c < 1$  measuring that strength. The index can possibly be normalised between 0 and 1, by computing an absolute maximum for  $\varphi_c$  which would result from the free inverse (maximising) problem, but this has not been done in this paper.

### 3. Applications and comparison.

As said in the introduction, both methods have been applied to Dutch and Belgian regions, mostly provinces; several versions were explored, but only the main results will be taken up here.

#### 3.1. The Netherlands.

The connectropy results will be shown here without the directional links that created them; it should be remembered that they were obtained without considering connectivity of the spatial graphs generated, i.e. the *local* contiguity of all the regions concerned.

Table 1 first presents the connectropy results.

*Table 1: connectropy groupings for the Netherlands.*

1. Overijssel, Utrecht, Gelderland, Limburg;
2. Noord-Holland, Noord-Brabant, Zuid-Holland;
3. Groningen, Zeeland, Flevoland, Drenthe, Friesland.

The following groupings (tables 2 and 3) emerged from the cluster analysis, the number of clusters having been set at four, the usual grouping of Dutch provinces; computations were made without and with contiguity constraints.

*Table 2: cluster groupings for the Netherlands without contiguity constraints.*

1. Flevoland, Drenthe, Zeeland, Friesland, Groningen;
2. Overijssel, Limburg, Utrecht, Gelderland;
3. Noord-Brabant, Noord-Holland;
4. Zuid-Holland.

There exists a remarkable matching between the two groupings; apart from the fact that Zeeland, a south-western province, does not fit into group 3 of table 1 and into group 1 of table 2, and that the most active province, Zuid-Holland, has been isolated (due to a four-cluster grouping instead of a three-cluster one), one obtains three macro-regions that satisfy in fact the contiguity constraints (see the map of the Netherlands in the appendix, section 6): Center-East (respectively groups 1 and 2 from the two tables), South-West (groups 2 and 3-4) and North-East (groups 3 and 1, with Zeeland the exception, as noted above).

Cluster analysis with contiguity constraints gave the following results (table 3).

*Table 3: cluster groupings for the Netherlands with contiguity constraints.*

1. Drenthe, Friesland, Flevoland, Groningen, Overijssel;
2. Utrecht, Gelderland, Limburg;
3. Zeeland;
4. Noord-Brabant, Noord-Holland, Zuid-Holland.

The main difference with tables 1 and 2 is the fact that Overijssel has switched to the North-Eastern group, while Zeeland has now been isolated.

It should be remarked that none of the groupings reproduces the so-called “Rimcity” (Noord-Holland, Zuid-Holland, Utrecht), and that the groupings show more variance in the cluster analysis than in the connectropy one.

### *3.2. Belgium.*

Again the main results only will be shown, along the lines of section 3.1.

Table 4 presents the connectropy outcomes.

*Table 4: connectropy groupings for Belgium.*

1. Bruxelles-Capitale, Antwerpen, Oost-Vlaanderen, Wesrt-Vlaanderen;
2. Limburg, Liège, Hainaut, Vlaams-Brabant;
3. Brabant Wallon, Namur, Luxembourg;
4. Extra-territorial Units.

The extra-territorial units are international organisations (embassies, NATO, EU). It should further be noted that, apart from Bruxelles-Capitale, the groupings are contiguous.

Tables 5 and 6 present the cluster analysis results, without and with contiguity constraints.

*Table 5: cluster groupings for Belgium without contiguity constraints.*

1. Extra-territorial Units, Luxembourg , Brabant Wallon, Namur;
2. Limburg, Vlaams-Brabant, Hainaut, Liège;
3. West-Vlaanderen, Oost-Vlaanderen, Bruxelles-Capitale;
4. Antwerpen.

Apart from the extra-territorial units and Bruxelles-Capitale, the groupings are contiguous; also notable is the isolated position of Antwerpen.

*Table 6: cluster groupings for Belgium with contiguity constraints.*

1. Extra-territorial Units;
2. Luxembourg, Brabant Wallon, Namur;
3. Limburg, Vlaams-Brabant, Hainaut, Liège, Bruxelles-Capitale;
4. West-Vlaanderen, Oost-Vlaanderen, Antwerpen.

The reshuffling between the three tables is more perturbing than was the case of the Netherlands; in particular the North-Center-South partition (the Belgian Political-regional organisation) does nowhere show up clearly.

#### *4. Conclusions.*

The analysis of section 3 has to be completed with other exercises; in particular, only one criterion has been used, to wit total GRP; outcomes would undoubtedly be different, if e.g. per capita GRP would have been used. But the methods exposed and applied could be used to better expose the inner workings of the regional-economic systems concerned.

#### *5. References.*

Kaashoek, J.F., Paelinck, J.H.P. and Zoller, H.G., 2004, On Connectropy, in A. Getis, J. Mur and H.G. Zoller (eds), *Spatial Econometrics and Spatial statistics*, Palgrave, Basingstoke, pp. ....

Paelinck, J.H.P., 2004, Spatial Econometrics and Clustering: On Regime Selection, *paper prepared for the annual ERSA Meetings*, Oporto, August 2004.

#### *6. Appendix: maps.*



