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

CHAPTER 6

Conclusions and further research

This thesis proposes new methodologies for designing regions from lower level territorial units (areas) considering not only their characteristics but also the relationships between them.

These methodologies obviate the need for *ad-hoc* regionalisation to obtain territorial units that are representative of the phenomenon under consideration. This aspect is especially important since statistical and econometric results are sensitive to different levels of aggregation and scale.

After a survey of the most frequently used regionalisation methods, in chapter 2 we present a linear optimisation model to find the optimal aggregation of different areas in a given number of regions, by considering a geographical contact matrix and a relationship matrix. The minimisation of the "internal" heterogeneity of each region helps to identify regions that are homogeneous according to the criteria considered.

The possibility of treating the regionalisation problem as a linear model ensures that, due to its mathematical properties, the feasible region is convex and, as a result, the optimal solution can be found. Other advantages of this kind of formulation are that it is easy to implement in a great variety of commercial software without paying a high price for it, and its flexibility when some changes or additional constraints are needed. ojo!

The obtained empirical evidence permits to affirm that the proposed methodology has a great capacity to identify different complex territorial configurations. The model takes into account the contiguity constraint but without conditioning the shapes that those regions can adopt.

It is also important to highlight that the model permits to easily introduce additional restrictions in the regionalisation process. As an example, it has been shown the possibility of introducing two additional restrictions: the minimum population requirement and the mandatory isolation.

The next stage deals with the thesis' second specific objective. In chapter 3 an algorithm called *RASS* (*Regionalisation Algorithm with Selective Search*) is formulated as a way of improving the computational capacity of the direct optimisation model. This algorithm uses the advantages of applying direct optimisation to a given territorial portion which varies in each iteration, thanks to a selective search strategy. These characteristics permit the *RASS* to escape from local optimum.

proof

The results obtained with the *RASS* demonstrate its utility, since the global optimum was found in all the simulations considered and in a considerably shorter running time than when the direct optimisation model is applied.

Table 6.1 shows the main characteristics of regionalisation models proposed in this thesis dissertation and the previous models. As can be seen, both the linear optimisation model and the *RASS* algorithm overcome some of the drawbacks inherent in existing methodologies.

A common characteristic in all the regionalisation methods presented is that the **number of regions** to be formed is an exogenous variable. These **regions are obtained automatically** by applying the proposed models. This is an advantage over regionalisation models based on clustering techniques, in which several trials must be run before the number of regions is determined.

To take into account the **relationships between areas** to be grouped, the proposed models incorporate them through a squared matrix that contains a relationship measure between each pair of areas. Cutting models only take into account relationships between contiguous areas ($w : E \rightarrow N$), and iterative reallocation algorithms as AZP does not use these relationships in their searching processes.

Table 6.1. Comparison between revised regionalisation models and the linear optimisation model propose in this thesis dissertation.

Regionalisation methodology / Characteristics	Clustering		Mathematical programming				Iterative algorithms	
	Conventional	Adapted	Non linear regionalisation	Cutting models	Based on centroids	**Linear optimisation model**	Iterative realocation algorithms	***RASS***
The number of groups is given	✓	✓	✓	✓	✓	✓	✓	✓
Automatic regionalisation	×	×	✓	✓	✓	✓	✓	✓
Relationships along areas (n×n)	✓	✓	✓	×*	✓	✓	✓**	✓
Non metric relationships	✓	✓	✓	✓	×	✓	✓**	✓
Contiguity relationships	×	✓	✓	✓	×	✓	✓	✓
Shapes flexibility	✓	✓	✓	✓	×	✓	✓	✓
Optimal solution	×	×	×	✓	✓	✓	×	×
Acceptable for large problems	×	×	×	×	×	×	✓	✓
Initial feasible solution	×	×	×	×	×	×	✓	✓

Shading columns: models proposed in this thesis dissertation.

* It is only taken into account the relationships between each area an its neighbouring areas (i.e. first order relationships).

** They can be incorporated into de objective function but they are not taken into account in any step of the algorithm.

Non metric relationships between areas can be used in the models proposed. In contrast, models based on the selection of centroids have to use metric relationships in order to ensure that the regions obtained are contiguous after the assignation process.

The **contiguity relationships** between areas to be grouped are an important input in the models proposed. This information is not taken into account in clustering models when the two-stage regionalisation strategy is applied. Centroid-based regionalisation models do not use the contiguity relationships because in the assignation process contiguity is obtained by using metric relationships between areas.

Shape flexibility is an important characteristic in some regionalisation processes where regional shapes need only depend on data characteristics and are not imposed by the methodology. Centroid-based regionalisation process tend to produce compact areas.

Finding the **global optimum** solution of a regionalisation problem can only be guaranteed by applying linear optimisation models such as cutting models, centroid models and the linear regionalisation model proposed in this thesis dissertation. In iterative models the optimal solution cannot be found, but these models are well suited to solving **large regionalisation problems**. Finally, only in iterative regionalisation models is an **initial feasible solution** necessary in order to start the search process.

Finally, two different regionalisation processes were applied in order to design analytical regions that are homogeneous in terms of the interest variable: the first one based on the K-means algorithm applying the two-stage regionalisation strategy and a second one by applying the *RASS* algorithm.

Both models were applied in the context of provincial unemployment rates in Spain to compare normative regions (NUTS) with the analytical regions obtained. The results show that more homogeneous regions were designed when applying analytical regionalisation tools. Two other interesting results are related to the fact that analytical regions were also more stable during the analysed period and to the effects of scale in the regionalisation process.

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 designing appropriate geographical units and related with the analysed phenomena.

Further research can be divided into theoretical and empirical lines:

Theoretical research lines:

1. To incorporate different objective functions suitable for other regionalisation problems. Some of these functions may be non linear.

2. To endogenize the number of regions, so that the regionalisation model chooses the most appropriate number of regions. In statistical literature some indicators have been proposed in order to define the number of groups to design, but these indicators have not been introduced into regionalisation models.

3. To explore the theoretical implications of regionalisation processes in the spatial econometric field. Is there any relation between analytical regionalisation processes and the presence of spatial autocorrelation?

Empirical research lines:

1. To formulate a new heuristic model that allows solution of larger regionalisation problems, without incorporating the linear optimisation problem. In this heuristic it will not be necessary to define an initial feasible solution and it will be possible that the number of regions m takes values between 1 and n. $1<m<n$. It is important to note that, as in the *RASS* algorithm, the relationships between areas will play an important role in the search process.

2. To apply the proposed regionalisation models in other contexts. A possibility is to apply these models in structural analysis cases where the final objective is to design regions. Functional analyses could also be performed in which the regions designed represent an intermediate step in an applied econometric analysis.

References

Aarts, E. and Lenstra, J. K. (1997): *Local search in combinatorial optimization.* Chichester, New York [etc.]: John Wiley and Sons.

Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1993): *Network flows: theory, algorithms, and applications*. Englewood Cliffs, N.J: Prentice Hall, cop.

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.

Amrhein, C. G. and Flowerdew, R. (1992): 'The effect of data aggregation on a Poisson regression model of Canadian migration', *Environment and Planning A*, 24, 1381-91.

Anderberg, M. R. (1973): *Cluster analysis for applications.* New York: Academic Press.

Aurenhammer, F. (1991): 'Voronoi diagrams - a survey of a fundamental geometric data structure', *ACM Computing Surveys*, 23, 345-405.

Battiti, R. and Tecchiolli, R. (1994): 'The reactive tabu search', *ORSA Journal on Computing*, 6 (2): 126-40.

Benabdallah, S. and Wright, J. R. (1992): 'Multiple subregion allocation models', *Journal of Urban Planning and Development*, 118 (1): 24-40.

Browdy, M. (1990): 'Simulatea Annealing - an improved computer model for political redistricting', *Yale Law and Policy Review*, 8, 163-79.

Calciu, M. (1996): 'Une méthode of classification sous contrainte of contiguïté en géo-marketing' Cahiers of recherche of l'IAE of Lille (96/5). Université des Sciences et Technologies of Lille.

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 15.

Crescenzi, P. and Kann, V. (2004): 'A compendium of NP optimization problems'. http://www.nada.kth.se/~viggo/wwwcompendium/

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.

Dantzing, G. B. and Ramser, J. H. (1959): 'The truch dispatching problem', *Management Science*, 6, 80-91.

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).

Ferligoj, A. and Batagelj, V. (1982): 'Clustering with relational constraint', *Psychometrika*, 47, 413-26.

Ferligoj, A. and Batagelj, V. (1983): 'Some types of clustering with relational constraints', *Psychometrika*, 48, 541-52.

Fisher, M. M. (1980): 'Regional taxonomy', *Regional Science and Urban Economics*, 10, 503-37.

Fotheringham, A. S. and Wong, D. W. S. (1991): 'The modifiable areal unit problem in multivariate statistical analysis', *Environment and Planning A*, 23, 1025-44.

Glover, F. (1977): 'Heuristic for integer programming using surrogate constraints', *Decision Sciences*, 8, 156-66.

Glover, F. (1989): 'Tabu search, part I', *ORSA Journal on Computing*, 1, 190-206.

Glover, F. (1990): 'Tabu search, part II', *ORSA Journal on Computing*, 2, 4-32.

Gordon, A. D. (1996): 'A survey of constrained classification', *Computational Statistics & Data Analysis*, 21, 17-29.

Gordon, A. D. (1999): *Classification*. 2nd edition, Boca Raton [etc.]: Chapman & Hall/CRC.

Gower, J. C. and Legendre, P. (1986): 'Metric and euclidean properties of dissimilarity coefficients', *Journal of Classification*, 3, 5-48.

Graham, R. R. and Hell, P. (1985): 'On the history of the minimum spanning tree problem', *Annals of the history of computing*, 7, 43-57.

Haining, R. P., Wise, S. M. and Ma, J. (1996): 'The design of a software system for interactive spatial statistical analysis linked to a GIS', *Computational Statistics*, 11, 449-66.

Hiriart, J. B., Oettli, W. and Stoer, J. (1983): *Optimization: Theory and Algorithms*. New York [etc.]: Marcel Dekker, cop.

Horn, M. E. T. (1995): 'Solution techniques for large regional partitioning problems', *Geographical Analysis*, 27, 230-48.

Jobson, J. D. (1991): *Applied multivariate data analysis: Categorical and multivariate methods*. New York [etc.]: Springer.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983): 'Optimization by simulated annealing', *Science*, 220, 671-80.

Kohonen, T. (1984): *Self-Organisation and Associative Memory*. Berlin [etc.]: Springer.

Laport, G. and Osman, I. H. (1995): 'Routing problems: A bibliography', *Annals of Operations Research*, 61, 227-62.

López-Bazo, E., del Barrio, T. and Artís, M. (2002): 'The regional distribution of Spanish unemployment: A spatial analysis', *Papers in Regional Science*, 81, 365-89.

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, 343-70.

Macmillan, B. and Pierce, T. (1994): 'Optimization modelling in GIS framework: the problem of political redistricting', in Fotheringham, S. and Rogerson, P. (eds.), *Spatial analysis and GIS*, London [etc.]: Taylor & Francis, pp 221-46.

Macmillan, W. (2001): 'Redistricting in a GIS environment: An optimization algorithm using switching-points', *Journal of Geographical Systems*, 3, 167-80.

Maravalle, M. and Simeone, B. (1995): 'A spanning tree heuristic for regional clustering', *Communications in Statistics - Theory and Methods*, 24 (3): 625-39.

Martin, D., Nolan, A. and Tranmer, M. (2001): 'The application of zone-design methodology in the 2001 UK Census', *Environment and Planning A*, 33, 1949-62.

Matula, D. W. and Sokal, R. R. (1980): 'Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane', *Geographical Analysis*, 12, 205-22.

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): 82-8.

Neves, M. C., Freitas, C. C. and Câmara, G. (2001): 'Mineraçâo of datos em grandes bancos of datos geographical.' Brasil: Instituto nacional of pesquisas espaciais. Ministério da ciência e tecnologia.

Ohsumi, N. (1984): 'Practical techniques for areal clustering', in Diday, E., Jambu, M., Lebart, L., Pagès, J. and Tomassone, R. (eds.), *Data analysis and informatics*, Vol III, North-Holland, Amsterdam, pp 247-58.

Openshaw, S. (1977): 'Algorithm3: a procedure to generate pseudo random aggregation of N zones into M zones where M is less than N', *Environment and Planning A*, 9, 1423-28.

Openshaw, S. (1984): The modifiable areal unit problem, *Concepts and Techniques in Modern Geography*, 38 (GeoBooks, Norwich).

Openshaw, S. (1992): 'Some suggestions concerning the development of artificial intelligence tools for spatial modelling and analysis in GIS', *The Annals of Regional Science*, 26, 35-51.

Openshaw, S., Alvanides, S. and Whalley, S. (1998): *Some further experiments with designing output areas for the 2001 UK census*, Centre for Computational Geography. School of Geography. University of Leeds, Leeds.
(http://www.geog.leeds.ac.uk/people/s.alvanides/papers/somefur4.html)

Openshaw, S. and Rao, L. (1995): 'Algorithms for reengineering 1991 census geography', *Environment and Planning A*, 27, 425-46.

Openshaw, S. and Taylor, P. J. (1981): 'The modifiable areal unit problem', in Wrigley, N. and Bennett, R. J. (eds.), *Quantitative Geography*, London, pp 60-70.

Openshaw, S. and Wymer, C. (1995): 'Classifying and regionalizing census data', in Openshaw, S. (eds.), *Census Users Handbook*, 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 Gaile, G. L. and Wilmott, C. J. (eds.), *Spatial statistics and models*, Reidel, Dordrecht, pp 55-79.

Theil, H. (1967). Economics and Information Theory, Chicago: Rand McNally and Company.

Toussaint, G. T. (1980): 'The relative neighbourhood graph of a finite planar set', *Pattern Recognition*, 12, 261-68.

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): 'Regionalisation tools for exploratory spatial analysis of health data', in Fisher, M. M. and Gentis, A. (eds.), *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling, and computational intelligence*, Berlin [etc.]: Springer, pp 83-100.

Departamento de Econometría Estadística y Economía Española
UNIVERSIDAD DE BARCELONA

DESIGN OF HOMOGENOUS TERRITORIAL UNITS.

A Methodological Proposal and Applications.

(RESUMEN EN CASTELLANO)

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

TABLA DE CONTENIDO

DESIGN OF HOMOGENOUS TERRITORIAL UNITS.

A Methodological Proposal and Applications.

(RESUMEN EN CASTELLANO)

1. Introducción y objetivos

La utilización de información estadística a nivel regional y urbano cada vez está siendo más utilizada para llevar a cabo análisis de carácter económico. Sin embargo, una de las principales cuestiones a resolver a la hora de realizar dicho tipo de estudios consiste en la selección de las unidades geográficas que serán utilizadas. Frente a esta cuestión los investigadores tienen dos alternativas: la primera consiste en la utilización de agrupaciones territoriales oficialmente establecidas a partir de criterios normativos (unidades geográficas como municipios, provincias, comunidades autónomas, etc.) y, como segunda alternativa, basar el estudio en unidades territoriales que estén directamente relacionadas con el fenómeno que se estudia (unidades analíticas). En este sentido, hay que destacar que frecuentemente las unidades analíticas no coinciden con las normativas, motivo que no impide que se lleve a cabo el estudio en cuestión a pesar de las consecuencias adversas que se pueden producir. De hecho, es conocido que la sensibilidad de los resultados a la agregación de datos geográficos puede traer consecuencias no deseables en el análisis (Openshaw, 1984), lo cual implica que dicha agregación no puede hacerse arbitrariamente. En este contexto, son numerosas las metodologías que se han desarrollado para el diseño de unidades territoriales que cumplan con unos requerimientos preestablecidos y que minimicen el peligro de obtener resultados erróneos, o poco representativos, como consecuencia del diseño de unidades territoriales inapropiadas. Sin embargo, dichas metodologías han sido formuladas para solucionar problemas de regionalización muy específicos, lo cual implica que en determinados ámbitos pueden resultar demasiado restrictivas o inapropiadas para cumplir con los requerimientos deseados.

Así pues, el objetivo de esta tesis consiste en implementar una nueva herramienta de regionalización para el diseño de unidades geográficas directamente relacionadas con el fenómeno analizado y que supere algunos de los inconvenientes de las metodologías actualmente disponibles.

Los objetivos específicos son:

a. Formular el problema de regionalización como un modelo de optimización lineal que permita tener en cuenta no sólo las características propias de cada unidad geográfica, sino también las relaciones de continuidad geográfica y relaciones de disimilaridad entre ellas.
b. Formular un heurístico capaz de resolver problemas de regionalización de mayor tamaño y que incorpore dentro de su estrategia de solución las características propias de un proceso de regionalización.
c. Comparar la homogeneidad de las regiones diseñadas con los modelos propuestos con los resultados obtenidos al utilizar agregaciones territoriales basadas en criterios normativos.

El resto de la tesis se organiza así: En el segundo capítulo se realiza una revisión de la literatura existente sobre diferentes metodologías para el diseño de unidades territoriales; en el tercer capítulo se presenta un modelo de optimización lineal para el diseño automatizado de regiones; en el cuarto capítulo se propone un heurístico para la solución de problemas de mayor dimensión; en el quinto capítulo se presenta una aplicación del heurístico en un entorno real y, por último, se presentan las principales conclusiones y futuras líneas de investigación.

2. Procesos de regionalización: Revisión de la literatura

Gran parte de las metodologías utilizadas en el campo de la regionalización utilizan técnicas enmarcadas dentro del análisis cluster. En este contexto, el problema de agregación de datos espaciales puede ser visto como un caso particular de cluster en el cual se debe asegurar la continuidad geográfica entre los elementos agrupados. Este caso especial de análisis cluster es llamado generalmente análisis cluster con restricción de continuidad espacial o regionalización[1]. Dentro de este tipo de análisis, destacan tres estrategias metodológicas: La agrupación en dos etapas, la inclusión de las coordenadas geográficas como variables de clasificación y la utilización de instrumentos adicionales para controlar la restricción de continuidad geográfica.

La agrupación en dos etapas consiste en aplicar, en una primera etapa, un método clásico de clustering, jerárquico o de particionamiento, utilizando únicamente los datos no-geográficos[2]. En la segunda etapa los grupos son revisados en términos de su continuidad geográfica, así, si a un cluster pertenecen áreas que son geográficamente discontinuas, entonces dichas áreas serán asumidas como regiones diferentes (Ohsumi, 1984). Entre las ventajas asociadas a dicha metodología Openshaw y Wymer (1995) resaltan el hecho de que en la primera etapa se garantiza la homogeneidad de las zonas creadas utilizando únicamente las variables de interés. Por otro lado, este método puede resultar útil como un medio para obtener evidencias de una dependencia espacial entre los elementos. Sin embargo, para los objetivos de la regionalización, puede convertirse en un inconveniente el hecho de que el número de grupos no dependa del investigador sino del grado de dependencia espacial[3].

---

[1] Un resumen de estas metodologías de agrupación se encuentra en Gordon (1999) y para los casos especiales de cluster restringido en Fisher (1980), Murtagh (1985) y Gordon (1996).

[2] Aquellos atributos propios de los elementos o de sus relaciones, por ejemplo, población, ingreso medio, número de hospitales, movilidad laboral entre municipios, etc.

[3] A mayor dependencia espacial habrá una tendencia hacia la creación de pocas regiones.

La segunda estrategia consiste en aplicar métodos clásicos de clustering pero forzando la continuidad geográfica por medio de la inclusión de las coordenadas geográficas de las áreas a agrupar, además de las variables de interés, en el cálculo de las similaridades o disimilaridades (Perruchet, 1983). Los principales inconvenientes derivados de la inclusión de variables geográficas como variables de clasificación son: la necesidad de que el componente geográfico tenga el peso suficiente para asegurar la continuidad geográfica (Wise *et al.*, 1997), y, dado que los elementos a agrupar son de tipo área, la solución puede variar en función del método utilizado para localizar el punto que representa cada una de ellas, sobretodo en aquellos casos donde las áreas son considerablemente grandes (Horn, 1995).

Por último, la más utilizada de las estrategias para la resolución de problemas de regionalización, consiste en controlar la restricción de continuidades geográficas utilizando instrumentos adicionales como la matriz de contactos o su correspondiente gráfico de continuidades. Dichos elementos son utilizados para realizar las modificaciones apropiadas a los algoritmos clásicos de clustering, jerárquicos o de particionamiento, asegurando el cumplimiento de la restricción de continuidad.

Para el caso de los modelos de particionamiento, los más utilizados en el contexto de la regionalización, destacan dos metodologías: la programación matemática y los algoritmos iterativos.

Con respecto a los métodos de programación matemática, Macmillan y Pierce (1994) definen el problema de regionalización como un problema de optimización combinatoria con el cual se determinará una agregación territorial óptima dado un número predeterminado de grupos a conformar. La solución propuesta por estos autores para asegurar la continuidad geográfica consiste en exponenciar la matriz de contactos teniendo en cuenta que para la formación de un grupo con n elementos continuos es necesario que la $(n-1)$ ésima potencia de dicha matriz no contenga elementos nulos. Esta solución implica que el espacio factible definido por las restricciones sea no convexo lo cual no asegura la obtención de un óptimo global.

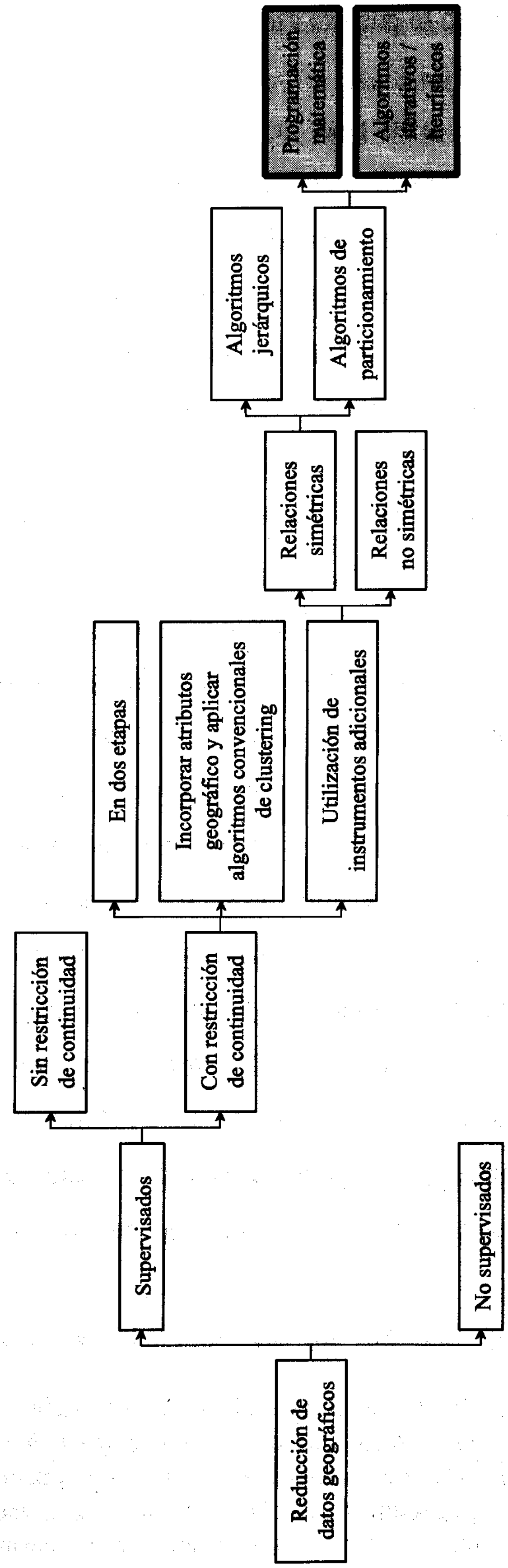
Dado que la solución de este tipo de problemas por métodos de optimización convencionales resulta extremadamente complejo, se han desarrollado otras metodologías en el campo de la regionalización que han resultado muy efectivas en aquellos casos donde el número de elementos a agrupar es muy grande. Dentro de este tipo de soluciones destacan distintos algoritmos como el "*automatic zoning problem*" (AZP) (Openshaw, 1977) que tienen como objetivo dividir un territorio en distintas áreas buscando además que dichas unidades territoriales minimicen, en lo posible, los efectos asociados al "Problema de la Unidad Espacial Modificable" (PUEM)[4].

Los modelos expuestos hasta ahora, se caracterizan por ser métodos supervisados, lo cual implica que el investigador conoce *a priori* la estructura de los datos que esta utilizando y que describen un fenómeno que se desea estudiar. Pero también existen métodos no controlados, útiles en casos donde el investigador desea analizar una gran cantidad de datos sobre los cuales no existe información de los factores que afectan el sistema. En estos casos surge la posibilidad de realizar un análisis no paramétrico que permita encontrar patrones y relaciones en un gran volumen de información. Una de las principales aplicaciones de estos métodos dentro del campo de la regionalización son los "*Self Organization Maps*" (SOM) propuesto por Kohonen (1984). Esta metodología, desarrollada en el campo de la inteligencia artificial ha generado opiniones divergentes entre los investigadores dada la falta de fundamentación teórica que dificulta la interpretación de los resultados (Openshaw, 1992).

La tabla 1 presenta un resumen de las metodologías descritas en esta sección.

[4] Openshaw planteó el problema de la unidad espacial modificable (PUEM), definiéndolo como una fuente potencial de error que puede afectar los resultados de aquellos estudios basados en información agregada geográfica, ya que estos pueden variar en función de la configuración de dicha agregación. Para más información ver Openshaw (1977), Openshaw y Taylor (1981), y en un contexto econométrico ver Fotheringham y Wong (1991) y Amrhein y Flowerdew (1992).

Tabla 1. Resumen de las metodologías para la reducción de datos geográficos



Fuente: Elaboración propia.

3. Modelo de optimización lineal para el diseño de unidades territoriales homogéneas

3.1. Descripción del modelo

Antes de su formalización matemática, se expondrán a continuación las características de dicho modelo y las hipótesis sobre las cuales fue construido.

El punto de partida de cualquier proceso de regionalización es determinar el territorio que se desea regionalizar, el cual debe estar formado por un número finito (n) de áreas geográficas de menor tamaño que forman un continuo geográfico $A=\{a_1, a_2, a_3,... a_n\}$. Este conjunto geográfico puede ser resumido en un grafo compuesto por n nodos representando cada una de las áreas y cuyos arcos representan las relaciones de continuidad entre éstas.

Para tal representación existen diferentes métodos de los cuales se ha seleccionado el más general, el "*Delaunay Triangulation*" (DT) (Aurenhammer, 1991), en el cual cada arco une aquellas áreas que comparten una frontera común.

El siguiente paso consiste en la incorporación de una serie de descriptivos sobre las relaciones entre cada una de las áreas (o nodos del grafo). Estas relaciones entre áreas son uno de los elementos más importantes dentro del proceso de regionalización propuesto, ya que su incorporación permite considerar interacciones entre las áreas para que su agrupación sea más coherente. Estas relaciones se incorporan en el modelo, a través de una matriz cuadrada y simétrica D_{ij} ($i = 1,2,...,n$ y $j = 1,2,...,n$), que contiene valores de disimilaridades entre cada par de áreas i, j. Dichas relaciones no tienen que ser métrica, es decir, no tienen que cumplir necesariamente con la propiedad de desigualdad triangular[5].

[5] Para más información, ver Gower y Legendre (1986).

Una vez se tienen los elementos descritos anteriormente (información sobre la configuración del territorio con el cual se trabaja y sobre las relaciones entre las áreas que lo forman), se puede iniciar el proceso de regionalización, que puede ser definido como la agrupación de las *n* áreas $\{a_1, a_2,...,a_n\}$ en *m* conjuntos no vacíos o regiones $\{1, 2,...,m\}$ tal que las áreas que pertenecen a cada región conformen un continuo geográfico.

Para definir dichas regiones será necesario seleccionar *n-m* arcos dentro del conjunto de arcos que conforman el grafo de continuidades. Estos *n-m* arcos son una condición necesaria pero no suficiente para formar cada una de las *m* regiones tal que cada región esté internamente conectada pero completamente desconectada de las demás regiones. Este sistema de configuración de regiones implica que el mínimo número de áreas en cada región será dos (un arco uniendo dos áreas), esto significa que $m = [n/2]$. Esta condición se hace menos restrictiva a medida que el número de áreas dentro del territorio sea mayor.

El último elemento en el proceso de regionalización consiste en definir un criterio de partición que permita detectar cual de las posibles configuraciones de *n* áreas en *m* regiones es la más adecuada. La medida de heterogeneidad seleccionada en esta tesis como criterio de agrupación se basa en el sumatorio de los elementos de la diagonal superior de la matriz de distancias (métricas o no métricas) entre las áreas que están dentro de una región. Conservando la notación utilizada por Gordon (1999), esta medida de heterogeneidad para la clase (o región) *r* (*Cr*) se expresa como:

$$H(C_r) \equiv \sum_{\{i,j \in C_r | i<j\}} d_{ij} \qquad (1)$$

Así pues, el objetivo para la construcción de *r* clases homogéneas puede ser planteado como la minimización del sumatorio de las medidas de heterogeneidad de cada clase *r*:

$$P(H,\Sigma) \equiv \sum\nolimits_{r=1}^{c} H(C_r) \tag{2}$$

3.2. Modelo matemático

Parámetros :

i, I índice y conjunto de áreas, $i = \{1,\dots,n\}$;

k, K índice y conjunto de regiones, $k = \{1,\dots,m\}$;

w_{ij} $\begin{cases} 1, \text{ sí } i \text{ y } j \text{ son continuas (comparten una frontera), con } i < j, \\ 0, \text{ en caso contrario;} \end{cases}$

M $Max\left(\sum_{j=1}^{n} w_{1j},\dots,\sum_{j=1}^{n} w_{nj}\right)$

N_i $\{j \mid w_{i,j} = 1\}$;

$D_{i,j}$ Medida de distancia entre áreas i y j, con $i < j$; la propiedad métrica no es importante

Variables de Decisión :

X_{ijk} $\begin{cases} 1, \text{ sí las áreas } i \text{ y } j \mid j \in N_i \text{ pertenecen a la misma región } k, \text{ con } i < j, \\ 0, \text{ en caso contrario;} \end{cases}$

Y_{ik} $\begin{cases} 1, \text{ si la área } i \text{ pertence a la región } k, \\ 0, \text{ en caso contrario;} \end{cases}$

T_{ij} $\begin{cases} 1, \text{ la medida de distancia entre } i \text{ y } j \text{ es considerada sí las dos áreas pertenecen a} \\ \quad \text{la misma región } k,\ i < j, \\ 0, \text{ en caso contrario;} \end{cases}$

Función Objetivo: $Min \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} \cdot T_{ij}$

Sujeto a :

$$T_{ij} \geq Y_{ik} + Y_{jk} - 1, \qquad \forall i, \forall j = 1,\dots,n \,;\ \forall k = 1,\dots,m \tag{3}$$

$$\sum_{i=1}^{n} Y_{ik} \geq 2, \qquad \forall k = 1,\dots,m \tag{4}$$

$$\sum_{k=1}^{m} Y_{ik} = 1, \qquad \forall i = 1,\dots,n \tag{5}$$

$$\sum_{j\in N_i} X_{ijk} \le Y_{ik} \cdot M, \qquad \forall i = 1,\dots,n \;;\; \forall k = 1,\dots,m \qquad (6)$$

$$\sum_{j\in N_i} X_{jik} \le Y_{ik} \cdot M, \qquad \forall i = 1,\dots,n \;;\; \forall k = 1,\dots,m \qquad (7)$$

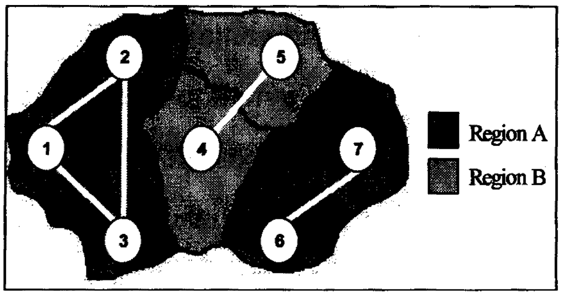
$$\sum_{i=1}^{n} \sum_{j\in N_i} X_{ijk} = \sum_{i=1}^{n} Y_{ik} - 1, \qquad \forall k = 1,\dots,m \qquad (8)$$

$$\sum_{i,j\in C} X_{ijk} \le |C| - 1, \quad \forall \text{subconjunto no vacio de } C \subseteq \{3,\dots,(n\text{-}2m+1)\}; \quad \forall k = 1,\dots,m \qquad (9)$$

$$X_{ijk} \in \{1,0\};\; Y_{ik} \in \{1,0\};\; T_{ij} \ge 0, \quad \forall i, \forall j = 1,\dots,n \;;\; \forall k = 1,\dots,m \qquad (10)$$

La función objetivo adoptada en el modelo propuesto busca la minimización del sumatorio de los elementos de la diagonal superior de la matriz de distancias entre los elementos que pertenecen a un mismo grupo (minimización de la heterogeneidad total). La restricción (3) permite seleccionar los elementos de la matriz de distancia que corresponden a las áreas que pertenecen a un mismo grupo. La restricción (4) impone que el mínimo número de áreas que conforman una región es dos. La restricción (5) impone que cada área debe ser asignada a una sola región. Las restricciones (6) y (7) imponen que sólo en el momento en el que una área i está asignada a una región k será posible asignar arcos de entrada o salida hacia las vecindades de dicha área ($j\in N_i$). La restricción (8) impone que el número de arcos necesarios para asegurar la continuidad de las áreas asignadas a una región debe ser igual al número de áreas asignadas a dicha región menos uno. Sin embargo esta restricción por si sola no asegura que la solución final genere regiones continuas, pues se pueden presentar casos como el que se ilustra en la figura 1, donde la región A, compuesta por las áreas 1, 2, 3, 6 y 7, satisface la restricción (8), pues para 5 áreas asignadas a dicha región fueron utilizados 4 arcos de unión, pero la existencia del ciclo de arcos 1-2, 1-3, 2-3 genera el rompimiento de la continuidad geográfica de la región. Por lo tanto será necesario controlar, además del número de arcos, la existencia de ciclos, punto que se tratará más detenidamente a continuación y del cual se deriva la restricción (9).

Figura 1. Configuración regional no factible



El problema de los ciclos ha sido tratado bajo el nombre de "*subtours*" en modelos de transporte como el "*Vehicle Routing Problem*" (VRP)[6] donde la configuración de una ruta, o "*tour*", para un vehículo no puede contener "*subtours*", y para controlar esta condición, el VRP incorpora la siguiente restricción:

$$\sum_{j,i \in S} X_{ijk} \leq |S| - 1, \ \forall \text{ subconjunto no vacío de } S \subseteq \{2,\ldots,n\};\ k=1,\ldots,m. \quad (11)$$

El inconveniente de este planteamiento es que el número de restricciones crece exponencialmente con *n* y *m*.

Sin embargo, la restricción del VRP propuesta a nivel teórico puede ser adaptada en el ámbito geográfico dado que es posible la identificación previa de los ciclos potenciales lo cual permite reducir considerablemente el número de elementos que pertenecen al conjunto *S*. Para la detección de ciclos potenciales basta con identificar aquellas combinaciones de arcos donde el número de arcos utilizados coincida con el número de áreas conectadas por dichos arcos. Este nuevo conjunto ha sido definido en la ecuación (9) como *C*.

Dada las características propias del proceso de regionalización que se propone, el nuevo conjunto *C* sólo tendrá que ser evaluado para potenciales ciclos que involucren un mínimo de 3 áreas y un máximo de (*n*-2*m*+1) áreas.

---

[6] Propuesto originalmente por Dantzing y Ramser (1959). Un resumen sobre modelos derivados del planteamiento original se encuentra en Laport y Osman (1995).

Así, la restricción (9) surge como una modificación de la definición del conjunto S propuesto en el VRP, lo cual trae como ventaja adicional una importante reducción del número de restricciones a satisfacer, evitando que el número de restricciones crezca exponencialmente con n y m. Esto hace posible la utilización de "*software*" comercial para la solución de problemas de regionalización con un mayor número de áreas y regiones.

Por último, en la restricción (10) del modelo, sólo es necesario fijar como variables binarias a X_{ijk} y a Y_{ik}. Aunque la variable T_{ij} se ha definido como positiva, no como binaria, siempre tomará valores de 0 o 1 por la combinación de la restricción (3) y el objetivo de minimización del modelo.

Finalmente, cabe destacar que dicho modelo puede ser fácilmente ampliado con el objetivo de satisfacer restricciones especificas que se deseen imponer a un problema de regionalización. Algunas restricciones adicionales implementadas en el modelo incluyen la imposición de que las regiones diseñadas tengan una población superior a un mínimo preestablecido y también la condición de que un determinado conjunto de áreas queden asignadas a regiones diferentes, restricción que puede ser útil, por ejemplo, a la hora de asignar mercados de influencia a unidades empresariales previamente establecidas.

3.3. Resultados computacionales

El modelo de optimización propuesto fue aplicado en diferentes ejemplos con el objetivo de conocer su capacidad para identificar diferentes tipos de configuraciones regionales. Los resultados obtenidos ponen de manifiesto la capacidad del modelo para diseñar regiones con variedad de formas, por ejemplo franjas litorales, pre-litorales e interior, o regiones en forma de coronas, entre otras.

En la tabla 2 se presentan los tiempos promedio de ejecución necesarios para alcanzar la solución óptima en problemas de regionalización con diferentes grados de complejidad. Dichos ejemplos fueron generados aleatóriamente, pero cuidando que las matrices de contactos generadas fueran realistas en términos de densidad y número de vecindades por área.

Tabla 2. Tiempo promedio de ejecución, en segundos, para diferentes combinaciones de áreas y regiones

		Regiones		
		2	**4**	**6**
Áreas	**5**	<1*	-	-
	8	<1*	3.00	-
	11	<1*	19.00	-
	14	5.80	117.40	2,571.00
	17	2.20	2,458.20	42,283.80

Nota: Cinco ejemplos para cada combinación de áreas y regiones.
* Tiempos de ejecución menores de un segundo.
Fuente: Elaboración propia.

A pesar de que la especificación de la restricción para la eliminación de ciclos (9) ha permitido reducir de forma significativa el número de restricciones a satisfacer, se puede observar, a medida que aumenta la complejidad del problema el tiempo de ejecución tiende a incrementar sustancialmente.

Como respuesta al problema de los tiempos de ejecución, en la próxima sección se presentará una alternativa metodológica, basada en modelos heurísticos, que permita solucionar problemas de regionalización más complejos.

4. Una solución para el problema computacional: "Regionalization Algorithm with Selective Search" (RASS)

4.1. Descripción del modelo

Una de las características que más interesa a la hora de aplicar los modelos matemáticos a problemas reales, los cuales involucran generalmente una gran cantidad de áreas, es el tiempo que tardan dichos modelo en encontrar la solución óptima. Con el objetivo de mejorar este tiempo, en la literatura se ha propuesto la utilización de diferentes heurísticos.

Los heurísticos de interés, teniendo en cuenta las características del modelo formulado, son aquellos que buscan la partición del territorio en un número preestablecido de regiones y dónde no se asigne un papel diferenciado a las áreas a ser agrupadas. Entre los heurísticos más utilizados en la literatura se encuentra el "Automatic Zoning Procedure (AZP)". Este heurístico, propuesto por Openshaw (1977), está basado en un procedimiento iterativo de prueba y error. Consiste en la optimización de una función objetivo *F(Z)*, donde *Z* es la asignación de cada una de las *N* zonas a una de las *M* regiones, cada zona debe ser asignada a un sola región y cada región debe tener como mínimo una zona. Entre las principales ventajas atribuidas a este heurístico destaca la posibilidad de utilizar cualquier función objetivo sensible a la agregación de zonas. Esta característica resulta de gran utilidad para tener una aproximación a los efectos límite (la deferencia entre maximizar y minimizar su valor) de la agregación de cualquier estadístico o modelo. También fue muy útil para demostrar la existencia del PUEM. Por otro lado, los principales inconvenientes atribuidos a este heurístico son su procedimiento de búsqueda local (restringida a la región seleccionada) y la fuerte dependencia de los resultados al punto de partida seleccionado. Por último, la estrategia de descartar la posibilidad de mover una zona que genere un descenso en la función objetivo puede hacer que el heurístico quede atrapado con relativa facilidad en óptimos locales[7].

Así pues, con el objetivo de implementar una estrategia de regionalización que evite en lo posible quedar atrapado en óptimos locales, el algoritmo *RASS* (*Regionalization Algorithm with Selective Search*) tiene como principal característica la capacidad de dirigir el proceso de búsqueda de forma selectiva, basándose en la información disponible sobre las relaciones entre áreas. El *RASS* incorpora dentro de su algoritmo el modelo de optimización presentado anteriormente en este mismo apartado y para conseguir mejoras locales de la función objetivo, consta de los siguientes pasos:

Paso 1: Partir de una solución factible de m regiones que agrupen n áreas.

Paso 2: Seleccionar de las m regiones el continuo geográfico más heterogéneo compuesto por r regiones, con $2 \leq r \leq (m-1)$.

$$H(C_m) \equiv \sum_{\{i,j \in C_k | i<j\}} d_{ij} \rightarrow Max\left(\sum_{m \in M_i} H(C_m) \right) \tag{12}$$

donde M_i es el conjunto formado por las diferentes alternativas de selección de r regiones continuas de las m disponibles.

Paso 3: Aplicar a las áreas contenidas dentro de las r regiones seleccionadas el modelo de optimización lineal propuesto para crear r^* regiones.

Paso 4: Seleccionar la región que entra (e): De las (m-r) regiones que quedaron fuera, identificar aquellas que estén en las vecindades del territorio formado por r^* regiones y seleccionar la región más similar a alguna de las r^* regiones que están dentro.

---

[7] Este último se ha intentado solucionar en posteriores propuestas como el "Simulated Annealing Variant of AZP (AZP-SA)" propuesto por Openshaw, el "*ANNEAL redistricting problem*" propuesto por Macmillan y Pierce (1994), el "*Tabu Search Algorithm*" (AZP-TABU) adaptado por Openshaw para problemas de regionalización ,y el heurístico basado en árboles de recorrido para las agrupaciones territoriales, como MIDAS, propuesto por Maravalle y Simeone (1995).

$$I(C_{d,f}) \equiv \sum_{i \in C_f} \sum_{j \in C_d | j>i} d_{ij} \rightarrow Min\left(I(C_{d,f})\right) \tag{13}$$

Dónde d es el conjunto formado por las r^* regiones que están dentro, y donde f es un subconjunto de las regiones que están fuera, formado por aquellas regiones que se encuentran en las vecindades de d. Cada una de las (m-r) regiones que quedaron fuera en el paso 2 sólo podrán ser seleccionadas como regiones entrantes una vez por ciclo (pasos 2 a 8).

Paso 5: Selección de la región que sale (s): Sale la región más diferente a la región que fue seleccionada para entrar (e) en al paso 4. La región que sale no puede romper la continuidad geográfica de d.

$$I(C_{d,e}) \equiv \sum_{i \in C_e} \sum_{j \in C_d | j>i} d_{ij} \rightarrow Max\left(I(C_{d,e})\right) \tag{14}$$

Paso 6: Incorporar al conjunto de r^* regiones la región seleccionada para entrar y extraer la región seleccionada para salir d=(d+e-s). A esta nueva configuración de r regiones aplicar el modelo de optimización propuesto para crear r^* regiones.

Paso 7: Repetir 4 a 6 hasta que las (m-r) regiones que quedaron afuera en el paso 2 hayan pasado por d, o hasta que no hayan regiones candidatas a entrar en las vecindades de d.

Paso 8: Calcular la función objetivo.

Paso 9: Si hay mejora en la función objetivo volver al paso 2. Si la función objetivo no mejora volver al paso 2 y seleccionar el siguiente grupo más heterogéneo. Repetir los pasos 2 al 8 hasta que no se encuentre una mejora significativa de la función objetivo a lo largo de un número C de ciclos o hasta que se haya agotado la lista de alternativas de selección de r regiones contiguas.

Cabe destacar que en el algoritmo *RASS* la aplicación de optimización directa a un grupo de regiones, en los pasos 3 y 6, permite lograr mejoras en la función objetivo que pueden generar cambios importantes en las configuraciones regionales, por la reasignación de un gran número de áreas entre las regiones. Además, los criterios utilizados en el paso 2 para la selección de las *r* regiones, junto con los criterios de entrada y salida de regiones de los pasos 4 y 5, buscan mantener dentro del modelo de optimización, paso 3, aquellas regiones que tengan el mayor potencial de mejorar la función objetivo tras su reconfiguración. De esta manera, se busca que la región que entra sea aquella que presenta la mayor posibilidad de contener áreas que óptimamente pertenezcan a otra región. Esta potencial reasignación de áreas se identifica asumiendo que dos regiones con áreas intercambiadas (en uno o ambos sentidos), aumenta la similitud entre dichas regiones. Por último, una vez identificada el área entrante, el criterio de salida (para mantener un número apropiado de áreas dentro del modelo de optimización) establece que la región que sale es la que, en los términos de disimilaridad utilizados, está más alejada del área entrante, y que por tanto dicha región tiene una menor posibilidad de intercambiar áreas con la región entrante. Otro aspecto a destacar es que *RASS* contiene, además de las condiciones de entrada y salida comentadas en el punto anterior y de la utilización de un método de optimización directo, dos condiciones cuyo objetivo es el de escapar de óptimos locales. En primer lugar, en cada ciclo todas las regiones que están afuera deben entrar una vez: para asegurar que todas las áreas sean evaluadas dentro del modelo de optimización, y, en segundo lugar, el retorno al paso 1 del algoritmo cada vez que termina un ciclo: para que la solución final no dependa de la selección inicial de las *r* regiones que entrarán dentro del modelo de optimización y para que la condición 1 no mantenga separadas dos regiones más tiempo de lo necesario.

4.2. Resultados computacionales y comparación con el método de optimización directa[8]

En la tabla 3 se presentan los tiempos de ejecución requeridos por el *RASS* para solucionar los ejemplos presentados en el apartado 3.3, Dichos tiempos son comparados con los tiempos requeridos por el modelo de optimización directa.

Los resultados obtenidos por el *RASS* alcanzan la solución óptima en el 100% de los ejemplos y en un tiempo sustancialmente menor que los requeridos por el modelo de optimización directa.

La última columna de la tabla muestra cómo después del primer ciclo el *RASS* la función objetivo disminuye en un 80% de la disminución total requerida para alcanzar el óptimo global.

Tabla 3 Comparación del RASS con el modelo de optimización directa.

Regiones	Áreas	Óptimos/5	Segundos (*RASS*)	Segundos (Directo)	(FOI - FO1c) / (FOI - SO*)
4	8	5/5	3.40	3.00	76.45%
	11	5/5	5.80	19.00	86.70%
	14	5/5	29.00	117.40	74.31%
	17	5/5	247.20	2,458.20	69.46%
6	14	5/5	25.20	25,710.00	85.93%
	17	5/5	250.00	42,283.80	66.71%

FOI= Función objetivo inicial, FO1c= Función objetivo después del primer ciclo, SO= Solución óptima.*
Fuente: Elaboración propia.

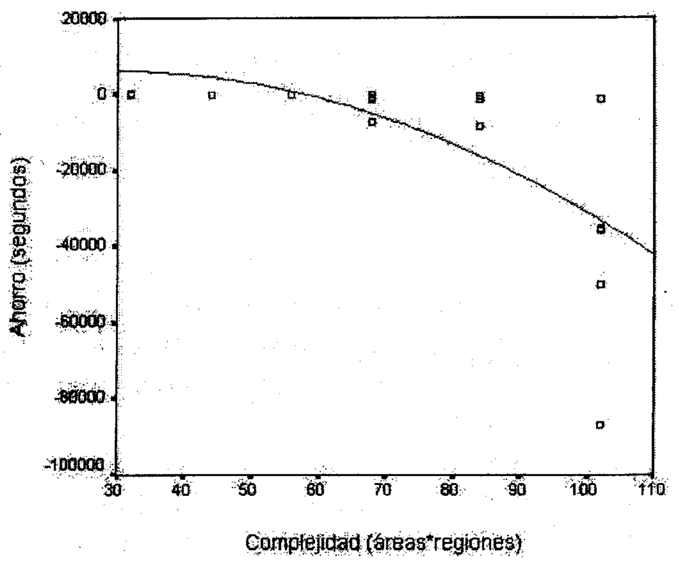
Utilizando la información disponible sobre los tiempos de ejecución requeridos por ambos métodos, optimización directa y *RASS*, es posible calcular para cada ejemplo el ahorro en tiempo conseguido con la aplicación del *RASS*, así, *AHORRO = tiempo*

[8] El software utilizado para ejecutar el modelo fue el Extended LINGO/PC 6.0 en un ordenador Pentium 4 a 2.40C GHz y 256 Mb de RAM. El heurístico fue implementado en Gauss 3.2.

RASS – tiempo optimización directa. La figura 2 muestra la relación entre dicha variable de ahorro y un indicador de complejidad que se ha calculado como el producto del número de áreas y el número de regiones a formar.

El resultado de esta relación indica que el algoritmo *RASS* se presenta como mejor opción en modelos con un grado de complejidad superior a 58. Dicho grado de complejidad se puede obtener con diferentes combinaciones de número de áreas y número de regiones.

Figura 2. Relación entre la complejidad del problema y el ahorro en tiempo al aplicar el *RASS*.



Fuente: Elaboración propia.

Así pues, los resultados obtenidos apuntan a que *RASS*, gracias a la incorporación de una rutina de optimización directa dentro de su algoritmo, tiene gran capacidad para alcanzar óptimos globales en un problema de regionalización. Sin embargo, es necesario tener en cuenta que la relación entre regiones a configurar (m) y áreas (n) debe ser establecida de tal forma que el número de regiones que entran al modelo de optimización (r) sea como mínimo de 2 y que dichas regiones contengan un total de áreas que se adapten a las capacidades computacionales del modelo. Para cumplir con dicho requerimiento se ha calculado que la relación m/n que se considera conveniente es $(m/n) \geq 14\%$.

5. Ilustración empírica de la metodología propuesta en el contexto del desempleo regional en España

Tal y como se ha comentado anteriormente, en el análisis regional aplicado, la información estadística está generalmente disponible a diferentes niveles de desagregación territorial con el objetivo de proveer información de interés a los posibles usuarios. Así pues a la hora de explotar dicha información existen dos alternativas. la primera consiste en utilizar las *regiones normativas* (municipios, provincias, etc.), o , en segundo lugar, diseñar *regiones analíticas* directamente relacionadas con el fenómeno a estudiar.

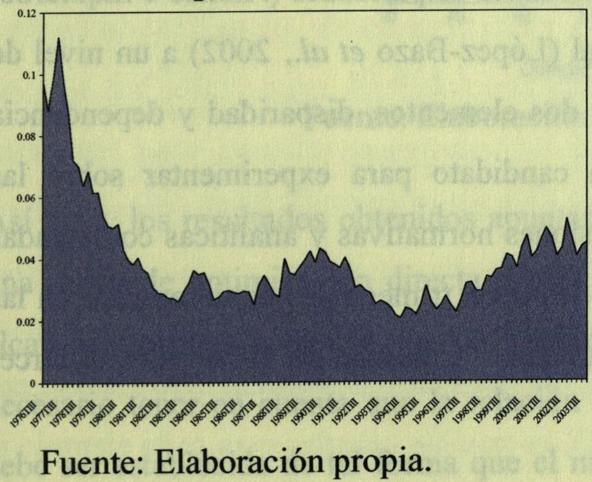
Existen muchas variables económicas cuyo análisis a un elevado nivel de agregación, por ejemplo nacional, resultan ser poco representativas debido a un elevado nivel de disparidad a niveles superiores de desagregación regional. Dichas disparidades regionales hacen necesaria complementar el análisis agregado con un análisis aplicado a un menor nivel de agregación con el objetivo de tener un mayor entendimiento del fenómeno a estudiar. Un claro ejemplo de esta situación se puede encontrar en el análisis de las tasas de desempleo en España. Estudios previos han demostrado que las tasas de desempleo en España presentan importantes disparidades (Alonso e Izquierdo, 1999), acompañadas de dependencia espacial (López-Bazo *et al.,* 2002) a un nivel de desagregación provincial (NUTS III). Estos dos elementos, disparidad y dependencia espacial, hacen de esta variable un buen candidato para experimentar sobre las diferencias que pueden existir entre las divisiones normativas y analíticas comentadas anteriormente. Pata tal análisis serán usadas las tasas trimestrales de desempleo en las provincias de la España peninsular, desde el tercer trimestre de 1976 hasta el tercer trimestre de 2003.

5.1. Desempleo regional en España: diferencias y dependencia espaciales

Con respecto a las disparidades espaciales, la figura 3 presenta los coeficientes de variación de las tasas de desempleo a un nivel de desagregación NUTS III para el período analizado. Como se puede observar, a lo largo del período, existe un importante grado de dispersión entre de las tasas de desempleo en las provincias españolas, con un valor promedio para el período de 43.03%. Esta dispersión es considerablemente superior durante la segunda mitad de la década de los 70. Por último, dicha dispersión se hace más evidente si se tiene en cuenta que la diferencia entre la mayor y menor tasa de desempleo es de 25.59%.

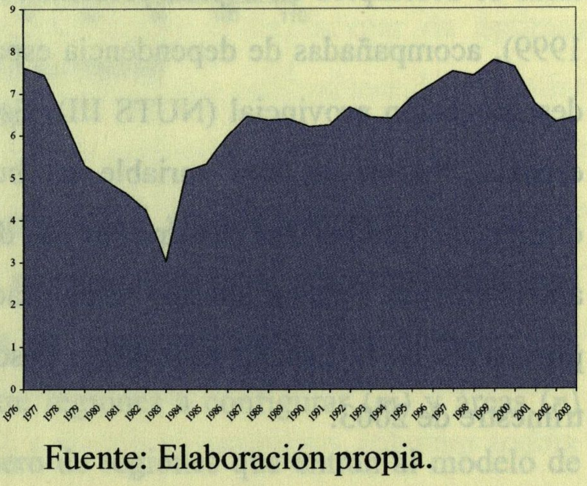
Con respecto a la dependencia espacial, la figura 4 presenta los valores estandarizados para el estadístico de la *I* de Moran (Moran, 1948) para un nivel de autocorrelación espacial de primer orden. Los resultados confirman la existencia de autocorrelación espacial de primer orden entre las provincias españolas a lo largo del período analizado.

Figura 3. Coef. de variación de las tasas de desempleo al nivel NUTS III.



Fuente: Elaboración propia.

Figura 4. *Z* de Moran de las tasas de desempleo al nivel NUTS III.



Fuente: Elaboración propia.

Así pues, se confirma el resultado ya apuntado en la literatura sobre la existencia de disparidades y de dependencia espacial, haciendo de dicha variable un buen candidato para realizar un proceso de regionalización, pues las diferencias espaciales dan lugar a la creación de grupos, mientras que la dependencia espacial justifica el hecho de que dichos grupos tiendan a ser geográficamente continuos.

5.2. Regiones normativas: clasificación NUTS.

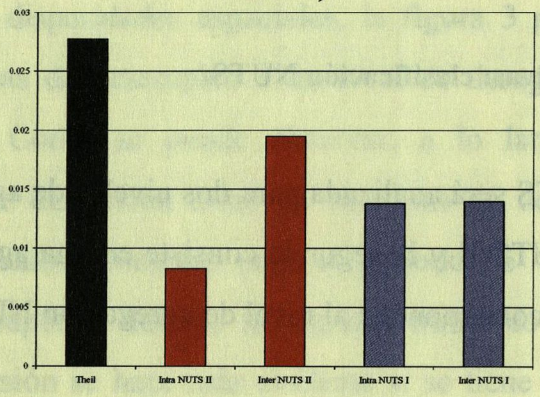
La división normativa NUTS será analizada para dos niveles de agregación: la primera consiste en 15 regiones NUTS II y la segunda consiste en una agrupación de estas 15 regiones en 6 regiones que corresponden al nivel de agregación NUTS I.

Para comparar el grado de homogeneidad de las regiones obtenidas para estos niveles de agregación, se ha calculado el índice de desigualdad de Theil (Theil, 1967). Dicho índice permite separar la desigualdad total en un componente que mide la desigualdad dentro de los grupos y otro componente que mide la desigualdad entre los grupos.

La figura 5 muestra el valor del índice de Theil y su desagregación en sus dos componentes cuando las tasas de desempleo provincial (NUTS III) son agregadas a niveles NUTS II y NUTS I. Como se puede observar, el grado de desigualdad interna es muy elevado, en términos relativos, en ambos casos, pero especialmente elevado para el nivel de agregación NUTS I.

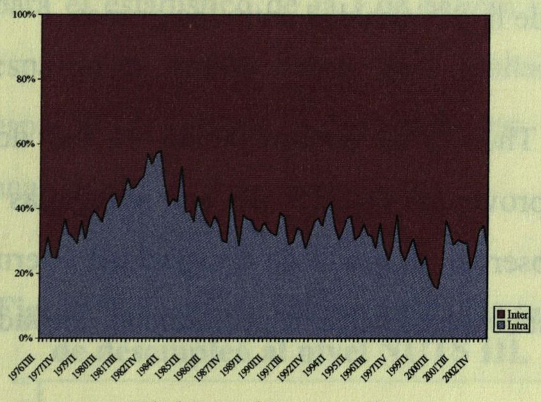
Uno de los objetivos de las unidades NUTS es que dichas unidades minimicen el impacto de los cambios de las estructuras regionales a través del tiempo. En las figuras 6 y 7 se presenta la desagregación del índice de Theil para cada uno de los trimestres disponibles, los resultados ponen en evidencia el comportamiento irregular del componente de desigualdad interna (Inter.). Dicha desigual alcanza su mayor nivel a principios de los ochenta y sus niveles inferiores durante el año 2000. También se nota como la proporción de desigualdad interna es mucho mayor en NUTS I que en NUTS II, esto se debe, en parte, al incremento en el nivel de agregación (de 15 a 6 regiones) y por la exigencia de que dicha agregación sea anidada.

Figura 5. Descomposición del índice de Theil para la tasa promedio de desempleo 1976 – 2003. (agregación NUTS III en regiones NUTS II y NUTS I).



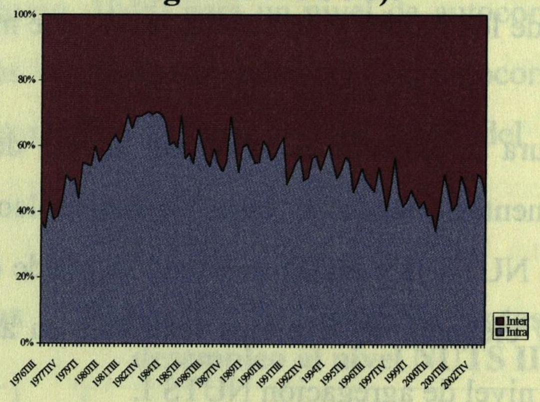
Fuente: Elaboración propia.

Figura 6. Descomposición del índice de Theil (agregación NUTS III en regiones NUTS II).



Fuente: Elaboración propia.

Figura 7. Descomposición del índice de Theil (agregación NUTS III en regiones NUTS I).



Fuente: Elaboración propia.

5.3. Regiones normativas vs. regiones analíticas

En este apartado se busca comparar, en términos de homogeneidad regional, los resultados obtenidos al aplicar métodos de regionalización analítica con las divisiones oficiales basadas en criterios normativos (NUTS).

Dos métodos de regionalización analítica serán utilizados: el algoritmo K-medias aplicado en dos etapas y el algoritmo *RASS*. Ambos métodos se aplicarán para agregar las 47 provincias en 15 regiones que puedan ser comparadas con el nivel de agregación

NUTS II y posteriormente las 15 regiones resultantes serán agregadas en 6 regiones para que puedan ser comparadas con el nivel de agregación NUTS I.

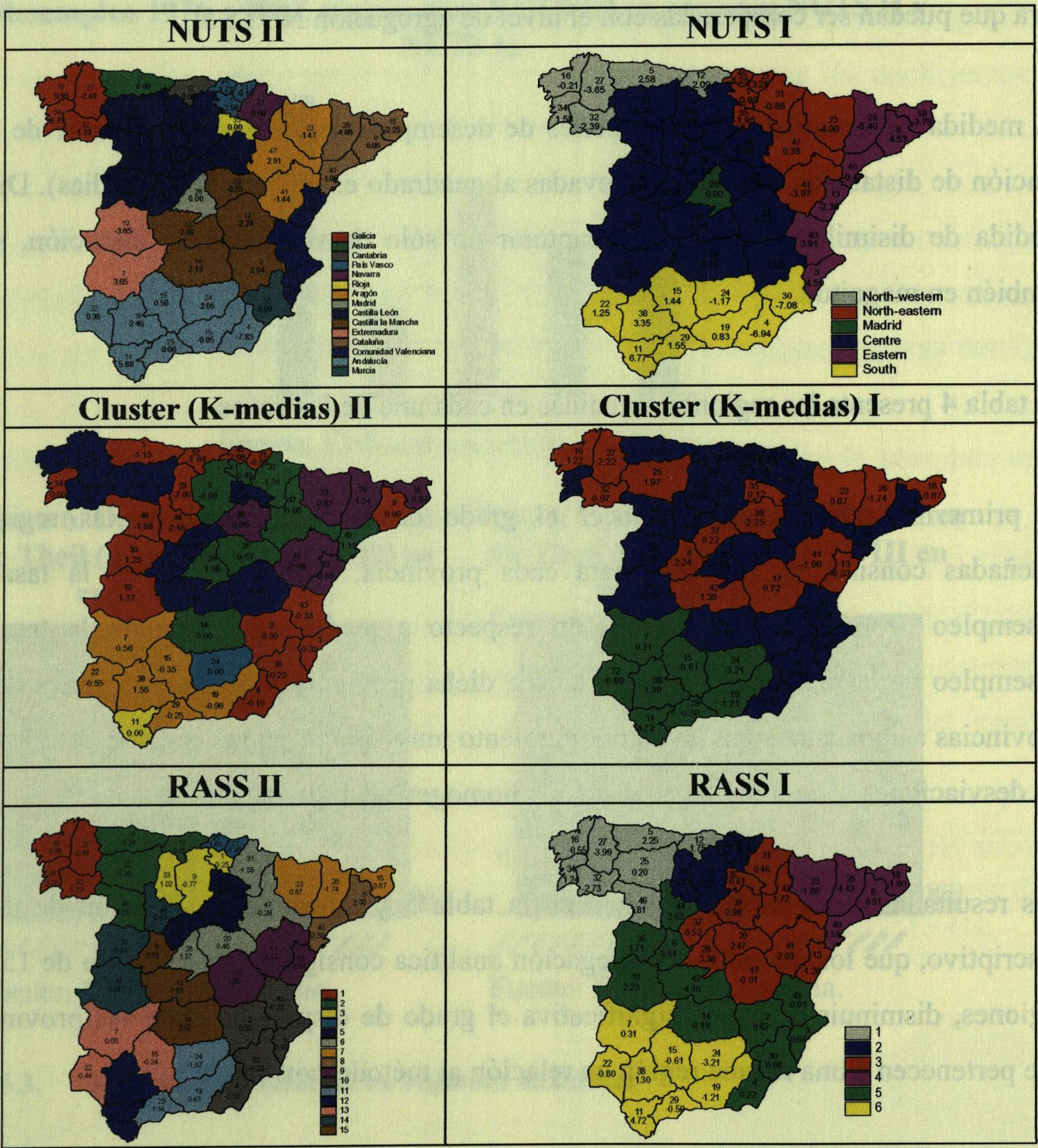
La medida de disimilaridad entre series de desempleo será calculada a partir de una función de distancias euclidiana (elevadas al cuadrado en el caso de K-medias). Dicha medida de disimilaridad permitirá capturar no sólo desviaciones en dirección, sino también en magnitud.

La tabla 4 presenta las regiones obtenidas en cada uno de los casos.

El primer descriptivo para conocer el grado de homogeneidad de las regiones diseñadas consiste en calcular, para cada provincia, la desviación de la tasa de desempleo promedio (1976-2003) con respecto a media aritmética de la tasa de desempleo de la región a la cual pertenece dicha provincia. Así, para regiones cuyas provincias asignadas tienen un comportamiento muy similar en las tasas de desempleo, las desviaciones serán muy bajas dado a la homogeneidad de las provincias.

Los resultados obtenidos se recogen en la tabla 5 y muestran, en función de dicho descriptivo, que los métodos de agregación analítica consiguen, para el caso de 15 y 6 regiones, disminuir de forma significativa el grado de dispersión entre las provincias que pertenecen a una misma región en relación al método normativo.

Tabla 4. Mapas de las agregaciones NUTS y de las agregaciones analíticas.



Fuente: Elaboración propia.

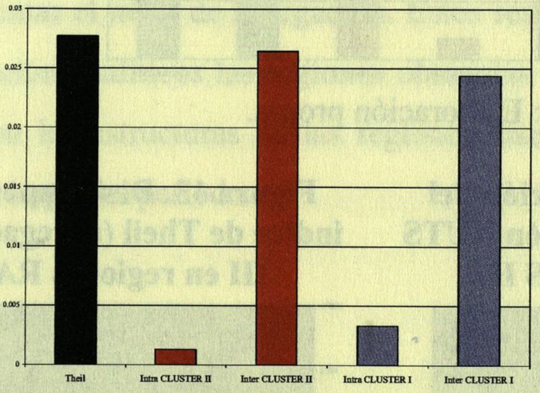
Tabla 5. Descriptivos sobre la desviación de la tasa promedio de desempleo en cada provincia con respecto al la región a la cual pertenece.

	NUTS II	RASS II	CLUSTER II	NUTS I	RASS I	CLUSTER I
Máximo	6.06	2.75	2.16	10.86	6.51	4.72
Mínimo	-7.83	-2.50	-2.22	-7.08	-4.42	-3.53
Rango	13.88	5.25	4.38	17.93	10.92	8.25
Desviación Estandar	1.90	0.74	0.69	2.30	1.49	1.21

Fuente: Elaboración propia.

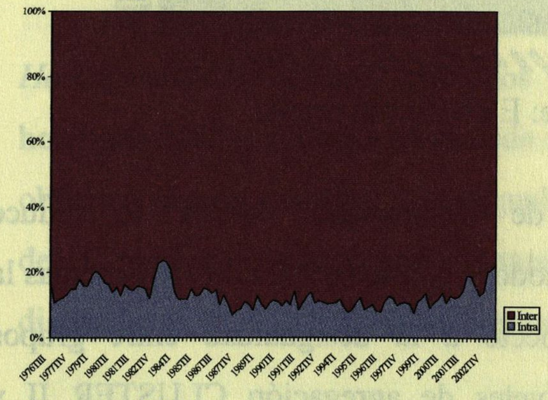
El segundo descriptivo utilizado consiste en calcular el índice de Theil, promedio y desagregado por trimestres, para las agrupaciones obtenidas con K-medias (figuras 8, 9 y 10) y *RASS* (figuras 11, 12 y 13). Estos podrán ser comparados con los índices presentados en el apartado anterior (figura 5).

Figura 8. Descomposición del índice de Theil para la tasa promedio de desempleo 1976 – 2003. (agregación NUTS III en regiones CLUSTER II y CLUSTER I).



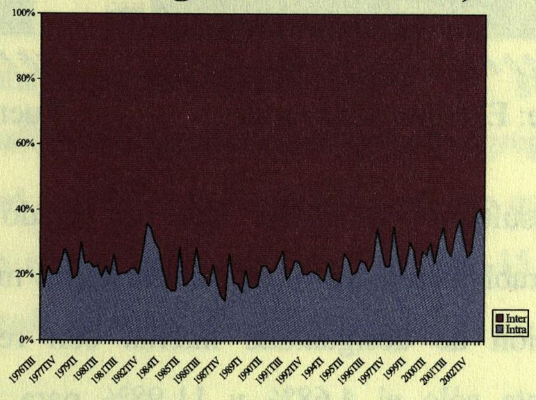
Fuente: Elaboración propia.

Figura 9. Descomposición del índice de Theil (agregación NUTS III en regiones CLUSTER II).



Fuente: Elaboración propia.

Figura 10. Descomposición del índice de Theil (agregación NUTS III en regiones CLUSTER I).



Fuente: Elaboración propia.