SUPERVISED REGIONALIZATION METHODS: A SURVEY*

By Juan Carlos Duque**, Raúl Ramos, Jordi Suriñach.***

**Department of Geography. Regional Analysis Laboratory (REGAL). San Diego State University. 5500 Campanile Drive, San Diego, CA 92182-4493. E-mail: jduque@rohan.sdsu.edu. Tel: 619 594 8032. Fax: 619 594 4938.

***Grup d'Anàlisi Quantitativa Regional (AQR), IREA, University of Barcelona, Dept. of Econometrics, Statistics and Spanish Economy, Avda. Diagonal 690, 08034 Barcelona, Spain;
email: rramos@ub.edu, jsurinach@ub.edu.

Abstract: This paper reviews almost four decades of contributions on the subject of supervised regionalization methods. These methods aggregate a set of areas into a predefined number of spatially contiguous regions while optimizing certain aggregation criteria. The authors present a taxonomic scheme that classifies a wide range of regionalization methods into eight groups, based on the strategy applied for satisfying the spatial contiguity constraint. The paper concludes by providing a qualitative comparison of these groups in terms of a set of certain characteristics, and by suggesting future lines of research for extending and improving these methods.

Keywords: regionalization, constrained clustering, analytical regions.

JEL codes: C21, R12, C61.

* The authors thank the anonymous IRSR reviewers for their insightful and helpful comments during the review process. The usual disclaimer applies. Jordi Suriñach and Raúl Ramos also acknowledge the support of CICYT SEJ2005-04348/ECON project.

1. Introduction

Statistical spatial data analysis often requires the aggregation of basic spatial units (areas) into larger units (regions) in order to preserve confidentiality, to minimize population differences, to reduce the effects of outliers or inaccuracies in the data, or simply, to facilitate the visualization and interpretation of information in maps (Wise et al., 1997, 2001).

An aggregation of this kind can be carried out in two different ways. The first uses as its reference point predefined official or normative aggregations, such as counties, districts or states. The second aggregates areas into analytical regions in such a way that the resulting regions are conveniently related to the phenomena under examination.

The Statistical Office of the European Communities (Eurostat, 2006) provides a clear differentiation between these two types of regions:

> *"Normative regions are the expression of a political will; their limits are fixed according to the tasks allocated to the territorial communities, to the sizes of population necessary to carry out these tasks efficiently and economically, or according to historical, cultural and other factors. Whereas analytical (or functional) regions are defined according to analytical requirements: functional regions are formed by zones grouped together using geographical criteria (e.g., altitude or type of soil) or/and using socio-economic criteria (e.g., homogeneity, complementarity or polarity of regional economies)."*

Although the use of normative regions has been a common practice among regional scientists, in many cases statistical inference based on regions of this type may be strongly affected by aggregation problems such as the ecological fallacy (Robinson, 1950), the modifiable areal unit problem (Openshaw, 1977a, 1977b, 1984; Openshaw and Taylor, 1981 and Arbia, 1989) or aggregation bias (Fotheringham and Wong, 1991; Amrhein and Flowerdew, 1992; Paelinck and Klaassen, 1979 and Paelinck, 2000).

In this paper we focus on methods that can be applied to design analytical regions. We will refer to them as "regionalization methods", to use a term that encompasses the wide range of fields in which these methods have been created or applied.[1] The variety makes it difficult to

[1] Regionalization methods are also known as spatially constrained aggregation methods, zone design and constrained clustering, among others.

provide a single definition of the regionalization problem, but we can enumerate some characteristics that are common to any method that can be used to define analytical regions:

1. All the methods aggregate geographical areas into a predefined number of regions while optimizing a particular aggregation criterion.
2. The areas within a region must be geographically connected (the spatial contiguity constraint).
3. The number of regions must be smaller than or equal to the number of areas.
4. Each area can be assigned to one and only one region.
5. Each region must contain at least one area.

Another characteristic of the methods considered here is that they are supervised, in that they assume prior knowledge about the aggregation process: relevant variables for the aggregation, number of regions to be designed, the regional spatial contiguity constraint and the existence of an aggregation criterion.

The purpose of this study is not to carry out a model competition but to make a qualitative comparison of the characteristics of the different methods that can be used in the context of the regionalization problem (as defined above). So our main contribution is to present a taxonomy of regionalization methods primarily based on the strategy used to satisfy the contiguity constraint, which is the main particularity of all the methods included in this paper. Previous related studies are the papers by Fischer (1980), Murtagh (1985) and Gordon (1996, 1999). However, these papers provide in-depth analyses of some of these methods rather than an exhaustive review.

The structure of the paper is depicted in figure 1, where the regionalization methods are classified into two main branches. The first branch, discussed in section 2, includes those approaches where the spatial contiguity constraint is indirectly satisfied. Section 3 covers the second branch of figure 1, which includes the approaches where the spatial contiguity constraint is explicitly included within the algorithm. In the last section, we present the main conclusions of this research.

FIGURE 1 ABOUT HERE.

2. Algorithms without an explicit spatial contiguity constraint

This section describes regionalization methods in which the contiguity constraint is indirectly satisfied, i.e. methods that do not incorporate an explicit procedure for satisfying the spatial contiguity constraint. This condition is usually applied *a posteriori*.

We classify these methods into two groups: Methods which apply conventional clustering and in which the resulting clusters are revised in terms of spatial contiguity, and methods which force spatial contiguity by maximizing regional compactness.

2.1. Regionalization via conventional clustering algorithms

This is probably the simplest regionalization method. Regionalization via conventional clustering algorithms was proposed by Openshaw (1973) as a methodological approach for regionalizing large datasets, comprising two stages. The first stage applies any conventional partitioning, or hierarchical, clustering algorithm to aggregate areas that are similar in terms of a set of variables.[2] In the second stage, each cluster is revised in terms of spatial contiguity by applying the following rule: If the areas included in the same cluster are geographically disconnected, then each subset of contiguous areas assigned to the same cluster is defined as a different region. Openshaw and Wymer (1995) formalized this method on a step-by-step basis for classifying and regionalizing census data.

Note that the number of clusters defined in the first stage is always smaller than or equal to the number of contiguous regions resulting in the second stage. Thus, adjustments in the number of clusters are required in order to obtain the number of regions desired. In some cases, this is not possible; for example, an increment (reduction) of one unit in the number of clusters in the first stage can generate an increment (reduction) greater than one in the number of regions in the second stage.

Openshaw and Wymer (1995) stressed the fact that regional homogeneity is guaranteed in the first stage. Moreover, this strategy may also help in providing evidence of spatial dependence between the areas. Thus, when the clusters in the first stage tend to be spatially contiguous, this may imply that the classification variables have some spatial pattern.

---

[2] See Johnson (1967), Anderberg (1973), Everitt (1993), Gordon (1999) for a review of clustering methods.

Another characteristic of this methodology is that it does not impose regional compactness.[3] In this case, the regional shape depends heavily on the spatial distribution of the classification variables and on the clustering algorithm chosen for the analysis. Johnston (1968), Lankford (1969) and Fischer (1980) pointed out that the selection of the clustering algorithm is very important for identifying certain spatial patterns. For example, the centroid and Ward's algorithms can easily identify circular and dense spatial patterns, whereas single linkage algorithm is useful to identify elongated spatial patterns.

Finally, when optimizing the aggregation criterion, conventional clustering algorithms like the k-means approach (MacQueen, 1967) only allow improving moves. This makes the algorithm converge quickly, mainly because it can easily be trapped in suboptimal solutions. It is also known that the final solution is very sensitive to changes in the initial centroids. One approach to this problem consists of solving it with different initial centroids and then select the best solution. Openshaw and Wymer (1995) proposed a simulated annealing[4] variant to this algorithm which allows non-improving moves as a way to force the algorithm to explore more solutions and avoid suboptimal ones.

2.2. Regionalization via maximization of regional compactness

Another way to obtain spatially contiguous regions is to force the regions to be as compact as possible. This strategy was introduced in the early 1960s by Weaver and Hess (1963) as a methodological approach to design political districts. The authors saw the opportunity to adapt the mathematical formulation for solving the warehouse location-allocation problem to the political districting problem. The aim is to select a subset of areas to be region centers (*warehouses*) to which the other areas (*customers*) are assigned.

The aggregation criterion consists of maximizing regional compactness by minimizing the sum of the "moments of inertia," defined as the product of the population per area and the squared distance from the centroid of each area to the region center it was assigned to.[5] It is important to note that region centers are not a decision variable but a parameter in the

[3] A restriction widely imposed in districting literature (Gillman, 2002).

[4] Simulated Annealing was formulated as an optimization procedure by Kirkpatrick et al. (1983) and first applied in the Redistricting Problem by Browdy (1990).

[5] This criterion makes highly populated areas good candidates for selection as the initial set of centers.

formulation. The only decision variable in the models is the assignation of areas to the predefined region centers.

The formulation also requires exactly equal populations in the regions. In order to satisfy this constraint fractional assignations of one or more areas to more than one region must be allowed. An iterative procedure fixes those fractional assignments and re-calculates the new regional centers. When the solution without fractional assignments leads to a change of the regional centers, the warehouse location-allocation model is solved again with this new set of centers. The process stops when no change of centers is needed after solving fractional assignments.

The satisfaction of the spatial contiguity constraint is not always guaranteed, since the assignation of the areas to their closest regional center is based on a weighted distance measure (population * distance). Thus, a final inspection of the final solution is required to correct spatial disconnections.

Hess et al. (1965) made a more formal presentation of this method. The fragmentation of areas is theoretically[6] solved by relaxing the equal population constraint to a "nearly equal" population requirement that allows the regional population to be between a lower and an upper bound. This relaxation makes it possible to formulate the problem as an integer programming model with a decision variable $x_{ij} = 1$ if the population of area j is assigned to the center i, and $x_{ij} = 0$ otherwise. $x_{ij} = 1$ with $i=j$, means that area i is selected as region center. A final revision for spatial contiguity is still required.

Kaiser (1966) proposed an aggregation criterion based on a weighted combination of two components. The first component is a measure of population equality, in which the population of each region should be as close as possible to the ratio between the total population and the number of regions to be designed. The second component is a measure of relative geometric compactness,[7] where the shape of each region should be as close as possible to a circle. This relative compactness is calculated as the proportion of the geometric moment of inertia for region j and the moment of inertia of a circle with the same geometric area. The minimum value of this quotient is one, signifying that region j is a perfect circle. For a given solution, the global

[6] It was theoretical because solving the new formulation was computationally difficult at that time. For empirical purposes the authors continued to use the procedure proposed by Weaver and Hess (1963).

[7] The use of a relative measure of compactness places more emphasis on regional shape than on regional size, since regional size depends on the spatial distribution of population when equal population is required. Thus, the lower the population density, the bigger the region must be in order to reach a given regional population.

compactness is measured as the average of the relative compactness for all regions. Kaiser's regionalization procedure starts from an initial feasible solution that is improved, in terms of the aggregation criterion, by moving areas between regions. Two types of moves are allowed: first, moving an area from its region to every other region, and second, exchanging every pair of areas belonging to different regions.[8] Only improving moves are accepted, which means that the process may well be trapped in local optimal solutions and be sensitive to the starting solution. The iteration process stops when no improving moves are possible. Finally, feasibility in terms of contiguity constraint depends on the weight given to the population equality component with respect to the compactness component.

Mills (1967) extends the Hess et al. (1965) location-allocation approach by taking into account natural boundaries in the regionalization process[9] in such a way that a region is not split by these types of boundaries. This condition is achieved by performing what he called "permanent assignations," which consist of assigning an area to a particular center in order to avoid this area being assigned to another center located on the opposite side of a given natural boundary.

Hess and Samuels (1971) extended the location-allocation methodology to the sales districting model. In this case, the regions are designed in order to equalize regional workload rather than regional population. The authors prefer the relaxed version proposed by Weaver and Hess (1963) where fractional assignations are adjusted *a posteriori.* The Hess and Samuels model is also extended by Zoltners (1979) to allow the incorporation of more than one attribute to be equalized across regions.

Helbig et al. (1972) formulated a heuristic and an integer programming model very similar to the one proposed by Hess et al. (1965). The authors assume that each area has a relatively equal population. This assumption makes it possible to replace the two sets of constraints related to the upper and lower bound for regional population by a set of constraints specifying the number of areas per region. Both the number of areas per region and the location of region centers are then adjusted iteratively by running the model several times. This modification avoids having to deal with fractional assignations of areas.

---

[8] Note that these moves do not take into account the feasibility in terms of spatial contiguity.

[9] In Mills' application, the natural boundary was the river Avon in Bristol.

George et al. (1997) incorporated modifications to Hess et al. (1965) and presented three new versions of the model. The first one incorporated nonlinear penalty functions for deviations from the target regional population, in such a way that the bigger the deviation from the desired regional population, the bigger the penalty. The second one offers the possibility of inducing the model to create a similar solution to a given pre-existing solution by introducing a function that reduces the distance from an area to a given center if they appear together in the pre-existing solution. The third one avoids having a region split by natural boundaries by introducing a penalty function that increases the distance between an area and the centers located in the opposite side of the natural boundary.[10]

Recently Bacao et al. (2005) proposed a methodology that applies genetic algorithms to define the location of the region centers. The algorithm starts by creating an initial set of solutions. Each solution comprises a set of region centers, the assignation of each area to the closest region centers, and a value for the aggregation criterion. With this initial set of solutions, new solutions are created by applying selection, crossover and mutation operators in order to improve the aggregation criteria. The algorithm stops when a predefined number of solutions are generated without improvements in the aggregation criteria.

Most of the methods derived from Weaver and Hess (1963) usually place the emphasis on regional equality, for example population or workload equality, and regional compactness. In these methods, regional homogeneity is assumed to be indirectly reached by imposing compactness. This assumption is related to Tobler's first law of geography: “Everything is related to everything else, but near things are more related than distant things” (Tobler 1970, p. 236). Another common characteristic of these approaches is that they incorporate the concept of regional center, which is usually located in a highly populated area within the region. In some cases, researchers are not interested in having this type of hierarchy in the regionalization process.

An alternative way of regionalizing via compactness, while taking into account regional homogeneity, is to include the *x* and *y* coordinates of the centroids of the areas as two additional classification variables. This can be seen as a two-objective aggregation criterion. The first

[10] This way of introducing information about natural boundaries is more relaxed than the permanent assignations proposed by Mills (1967). A similar strategy was implemented by Segal and Weinberger (1977), who replaced the squared Euclidean distance of a straight line between each center and all the other areas by the length of the shortest path between each center and all the other areas.

objective uses non-geographical classification variables (for example, socio-economic variables) to measure within-region homogeneity. The second objective uses geographical variables (for example, *x* and *y* coordinates) to measure the regional compactness. These two objectives are combined to calculate pairwise similarities between areas in such a way that areas geographically near to each other are likely to be assigned to the same cluster. Once the pairwise similarity has been estimated, either hierarchical or partitioning clustering algorithm may be used to aggregate the areas.[11]

There are many ways to combine geographical distances and non-geographical dissimilarities into a single pairwise similarity value. Webster and Burrough (1972), Cliff (1975), and Perruchet (1983) proposed different multiplicative and additive forms to combine such elements. These authors also applied hierarchical agglomerative algorithms for merging areas.[12]

Wise et al. (1997) formulated a multi-objective function with three objectives: a) homogeneity, measured as the within region variance of a set of attributes; b) compactness, measured as the distance between the centroid of each area and the centroid of the region it is assigned to; and c) equality, measured as the difference between the total value of an attribute for all areas in a region and the mean total of this attribute for all regions. These three objectives are then combined in an additive form and the areas are aggregated by applying a k-means clustering algorithm.

The main drawback in formulating a multi-objective aggregation criterion is that we have to determine the proper individual weight for each objective (homogeneity, equality, compactness, etc.). For example, the weight assigned to the compactness objective should be strong enough to guarantee spatial contiguity (Webster and Burrough, 1972; Wise et al., 1997 and Openshaw et al., 1998). This assignation of weights has been the main source of criticism of this methodology, since it becomes another source of variation in the final solution.

Regional compactness has been a very important characteristic in problems like political districting and sales territory assignment. But it may not be a desired characteristic when the

[11] Although this methodology may be considered a slight variation on the one introduced in section 2.1, there are two important differences that make us classify this strategy here. First, it achieves spatial contiguity by forcing the regions to be compact, which implies a different and more restrictive assumption about the shapes of the regions to design. Second, when using this methodology, the researcher has to deal with technical issues like how non-geographical and geographical variables are going to be combined into a single similarity function.

[12] Cliff et al. (1975) aggregated the areas by using a Minimum Spanning Tree algorithm, which is equivalent to a single linkage hierarchical agglomerative clustering.

researcher is interested in detecting spatial patterns that can lead to elongated regions. Note also that in these methods there is not much "freedom" in the way the relationships (similarities) between areas are defined. These methods require a similarity function directly related to the position of the areas in the geographical space; therefore, the closer the areas to a regional center, the higher the possibility of assigning those areas to that center. Finally, some of these methods represent the areas by their centroids. This representation is often criticized because the final solution is, in most cases, sensitive to the selection of those centroids. This sensitivity can become a big issue when the areas to be aggregated are large (Horn, 1995 and Martin et al., 2001).

3. Algorithms with an explicit spatial contiguity constraint

The methods covered in this section include, within their solution process, additional instruments that ensure the spatial contiguity of each region. This implies that these models require information about the spatial neighboring relationships between areas.

We classify these methods into three main categories: exact optimization models, heuristic models, and hybrid or mixed heuristic models.

3.1. Use of exact optimization models for the regionalization problem

The aggregation of *n* areas into *m* spatially contiguous regions while optimizing a predefined aggregation criterion is not an easy problem to solve optimally (Altman, 1998). Cliff and Haggett (1970), Cliff et al. (1975), and Keane (1975) used combinatorial mathematics to approximate the number of feasible solutions for a regionalization problem. Since this number depends not only on the number of areas and regions but also on how the areas are distributed in the space, these authors focused on providing the lower and upper bounds of the total number of aggregations. The upper bound results from a totally unconstrained formulation where each area can be grouped with any other area. The lower bound tries to incorporate the spatial distribution by assuming the extreme case where the areas form a chain, and each area can only be grouped with its neighbors.[13]

[13] The number of feasible solutions for a regionalization problem involving the aggregation of 10 areas into 5 regions is a number between 126 and 42,525 (Keane, 1975).

However, the impressive progress of computational capabilities has made it possible to provide optimal solutions for problems that only a few years ago were insurmountable. As an example, Bixby et al. (2000) uses a Patient Distribution System (PDS) problem with 49,944 rows, 177,628 columns and 393,657 nonzero variables to illustrate the significant improvements in running times through the different versions of CPLEX. With a 296 MHz Sun UltraSparc the running times went from 57,840 seconds with CPLEX 1.0 (1988) to 165 seconds with CPLEX 6.5 (1999).[14]

The main challenge for solving regionalization problems via exact methods is to find an efficient way to deal with the contiguity constraint. In this section, we describe different strategies for incorporating spatial contiguity constraint in optimization models.

Zoltners and Sinha (1983) improved the model proposed by Zoltners (1979). In this new proposal, the authors introduce information into the model about adjacency/contiguity between areas in order to control the contiguity constraint. There are many alternatives for defininge spatial contiguity between areas; in this particular case, Zoltners and Sinha use the road and highway network as a way to represent how areas are interconnected. This representation takes into account natural obstacles such as mountains, lakes, and rivers. Up to this point, the territory has been represented as a network with nodes representing the areas and links representing the roads connecting areas. Next, a sequence of steps is performed to reduce the number of links in the network. First, a set of areas are selected to be region centers, as many areas as there are regions to be designed. Second, the shortest path between each center and other areas is calculated using travel time as link weights. Finally, the solution from the shortest paths is used to obtain information on the number of adjacency levels between each area and the centers.[15]

With this information Zoltners and Sinha formulated a linear optimization model with a decision variable $x_{ij} = 1$ if region i includes area j, and $x_{ij} = 0$ otherwise. The key constraint that ensures spatial contiguity is that an area j can be assigned to center i at adjacency level k if there exists a neighboring area of j assigned to the same center at adjacency level k-1.

Duque (2004) formulated the regionalization problem as a Mixed Integer Programming (MIP) model. As in the method we just saw, Duque borrows concepts from graph theory in order to deal

[14] See Bixby (2002) for more information about this topic.

[15] In order to make the reduction less restrictive, the authors also included additional links taken from near optimal shortest paths solutions. This implies that there may exist more than one path connecting an area to a given center at different or equal adjacency levels.

with the spatial contiguity constraint. Thus, the areas and their neighboring relationships are represented as a connected graph with nodes representing areas and links representing first order spatial connectivity between areas.

According to graph theory, the aggregation of *n* areas into *m* contiguous regions can be achieved by selecting *n-m* links from the connected graph. These *n-m* links can be understood as a necessary but not sufficient condition. An additional condition for feasibility imposes that every pair of areas belonging to the same region must be connected by one and only one combination of links.[16] In the literature, this last condition has been referred to as the subtour breaking constraint.

Duque's formulation includes a set of subtour breaking constraints. It is clear that the enumeration of all potential subtours can be extremely difficult and inefficient. An alternative approach to dealing with subtours is proposed by Current et al. (1984) for the shortest covering path problem (SCPP). It consists of solving a relaxed formulation that does not take into account any subtour breaking constraints. Then, if the solution contains subtours, they are eliminated by adding the proper subtour breaking constraints and solving the problem again. This process is repeated until a feasible solution is obtained.

Three sets of binary decision variables are used in Duque's formulation:

$$X_{ijk} \quad \begin{cases} 1, \text{ if areas } i \text{ and } j \,|\, j \in N_i \text{ belong to the same region } k, \text{ with } i < j \text{ and } N_i = \text{ set of areas } j \text{ sharing} \\ \quad \text{a border with area } i, \\ 0, \text{ otherwise;} \end{cases}$$

$$Y_{ik} \quad \begin{cases} 1, \text{ if area } i \text{ belongs to region } k, \\ 0, \text{ otherwise;} \end{cases}$$

$$T_{ij} \quad \begin{cases} 1, \text{ if areas } i \text{ and } j \text{ belong to the same region } k, \text{ with } i < j, \\ 0, \text{ otherwise;} \end{cases}$$

The set of variables X_{ijk} are used to select the *n-m* links, and the set of variables T_{ij} are used to select all the pairwise dissimilarities between the areas assigned to the same region.

[16] For more information on the properties of this (and other) configurations, see Ahuja et al. (1993).

The linear objective function is formulated as the minimization of the sum of all pairwise dissimilarities between the areas assigned to the same region, $Min \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} \cdot T_{ij}$, with D_{ij} defined as the dissimilarity relationships between areas i and j, with $i<j$.

Later, Duque et al. (2006) formulated three new models for solving the regionalization problem. The three formulations have the same objective function used by Duque (2004), but each one applies a different strategy to satisfy the spatial contiguity constraint. They are named Tour^{RM}, Order^{RM}, and Flow^{RM}.

Tour^{RM} selects *n-m* links and prevents subtours by adapting the MTZ constraints with lifting proposed by Desrochers and Laporte (1991). MTZ constraints were originally formulated by Miller et al. (1960) for the Travelling Salesman Problem (TSP). The strategy consists of having a decision variable, u_i, that numerates the areas in such a way that area i must have a value lower than area j if there exists a directed link leading from i to j. Whereas the TSP imposes that each area can have at most one entering and one leaving link, Tour^{RM} allows areas to have more than one entering link in order to provide the flexibility needed to create any regional shape. The set of decision variables are:

X_{ij} $\begin{cases} 1, \text{ if the link from area } i \text{ to } j \mid j \in N_i \text{ is selected, with } N_i = \text{set of areas } j \text{ sharing a border with area } i, \\ 0, \text{ otherwise;} \end{cases}$

T_{ij} $\begin{cases} 1, \text{ if areas } i \text{ and } j \text{ belong to the same region, with } i < j, \\ 0, \text{ otherwise;} \end{cases}$

u_i Integer value assigned to area$i$

Order^{RM} is based on the contiguity constraint formulated by Cova and Church (2000). In this case, the spatial contiguity constraint is achieved by selecting a core area per region. The remaining areas are assigned to one core area by taking into account the spatial contiguity order between the area and its corresponding core area. Thus, area i can be assigned to core area j at order o if and only if there is at least one area in the neighborhood of i that is assigned to j at order o-1. The set of decision variables are:

X_{iko} $\begin{cases} 1, \text{ if area } i \text{ is assigned to area } j \text{ at order } o, \\ 0, \text{ otherwise;} \end{cases}$

T_{ij} $\begin{cases} 1, \text{ if areas } i \text{ and } j \text{ belong to the same region } k, \text{ with } i < j, \\ 0, \text{ otherwise;} \end{cases}$

Flow^{RM} solves the regionalization problem by adapting Shirabe's model for the spatial unit allocation problem (Shirabe, 2005). Shirabe's model is based on network flow theory in which the contiguity constraint is satisfied by designing a connected sub-network representing fluid movement from multiple sources to a single sink. Flow^{RM} extends Shirabe's model for a single region to solve the problem for m regions. The set of decision variables are:

F_{ijk} $\begin{cases} \text{Non - negative contiguous variable indicating the amount of flow} \\ \text{from area } i \text{ to } j \text{ in region } k; \end{cases}$

Y_{ik} $\begin{cases} 1, \text{ if area } i \text{ belongs to region } k, \\ 0, \text{ otherwise;} \end{cases}$

W_{ik} $\begin{cases} 1, \text{ if area } i \text{ is chosen as sink for region } k, \\ 0, \text{ otherwise;} \end{cases}$

T_{ij} $\begin{cases} 1, \text{ if areas } i \text{ and } j \text{ belong to the same region } k, \text{ with } i < j, \\ 0, \text{ otherwise;} \end{cases}$

Finally, we should mention two additional approaches to the regionalization problem in the context of optimization models. The first is a non-linear optimization model formulated by Macmillan and Pierce (1994) where, in order to guarantee that the N areas assigned to a region are spatially contiguous, it is sufficient to ensure that the (N-1)th power of their binary contiguity matrix has no zero terms. And second, Garfinkel and Nemhauser (1970) formulated a linear optimization model that requires the enumeration of all the feasible regions as input. Then, the binary decision variable x_j is used to select a predefined number of regions such that each area belongs to one and only one region. Mehrotra et al. (1998) used a similar approach, but in this case the set of feasible regions is generated with a column generation technique (Barnhart et al., 1998) which creates new solutions from an initial feasible solution.

Exact formulations for the regionalization problem are computationally intensive and their application is still limited to very small problems. As a very simple illustration, Table 1 presents the result for Tour^{RM}, Order^{RM}, and Flow^{RM} models using ILOG CPLEX 9.0 on a Sun Blade 2500, 500 MHz, 2Gb RAM to aggregate the fourteen counties in Vermont into three and nine regions.

TABLE 1 ABOUT HERE

3.2. Heuristic models for the regionalization problem

This section will cover different heuristic models for regionalizing. These models have been widely applied in the literature since the early 1960s, and they have proved to be highly effective in cases in which a large number of areas are to be aggregated. The main goal of heuristic models consists of finding, in a reasonable time, a solution as close as possible to the optimal, which in most cases is unknown. The ability to escape from local optimal solutions in order to reach a good solution is a very important component of the heuristic approaches. An additional challenge for heuristic models in the regionalization context is the ability to efficiently move from one solution to another neighboring solution without breaking the spatial contiguity constraint.

This review covers four types of heuristics: 1) heuristics based on hierarchical clustering algorithms where only contiguous regions can be merged, 2) heuristics where each region starts from an area (seed area) to which other neighboring areas are added, 3) heuristics that start from an initial feasible solution and search for improvements by swapping areas between regions, and 4) heuristics based on graph-theory.

3.2.1. Adapted hierarchical clustering algorithms

This methodology is another way of using conventional hierarchical clustering methods to solve regionalization problems. In this case an agglomerative hierarchical clustering algorithm is modified in such a way that only spatially connected regions (clusters) are allowed to be merged.

At the beginning of the algorithm, each area is a region by itself. Further iterations merge one region at a time until reaching a predefined number of regions. The research in this area explores the use of different methods of agglomerative hierarchical clustering, such as single linkage, complete linkage, average linkage and Ward's method, among others. Two characteristics are explored in these methods: first, their ability to identify different spatial patterns (Lankford, 1969

and Byfuglien and Nordgard, 1973), and second, how the spatial contiguity constraint affects the dendograms, which are a graphical representation of the clustering solutions at different scales (Ferligoj and Batagelj, 1982).

Other contributions to this topic have been made by Spence (1968), Webster and Borrough (1972), and Margules et al. (1985), who carry out comparisons between different agglomerative hierarchical techniques when imposing the contiguity constraint; Openshaw (1973), who proposes a way to speed up the regionalization procedure based on hierarchical aggregation techniques by reducing the number of calculations and restricting the size of the similarity matrix by retaining only the similarities between contiguous areas; and Perruchet (1983), who defines spatial contiguity based on the contiguity threshold, which is the maximum distance beyond which two areas are not contiguous.

Hierarchical approaches can be useful when the researcher is interested in nested solutions at different scales (number of regions). This nested solution occurs because when two areas are merged at a given scale they are forced to be together at higher scales. But, when the researcher is interested only in a specific scale this approach may not be the best option since the solution obtained is conditioned to the solution at every lower scale (Bunge, 1966).

3.2.2. Seeded Regions

The main characteristic of these heuristics is that each region is the result of selecting one area (seed area) to which other neighboring areas are assigned. The seeds can be generated one at the time, which implies that the next seed is selected after a region is completed; or in parallel, when all the seeds are selected at once.

This methodology was first proposed by Vickrey (1961) for solving districting problems. Vickrey's method starts by selecting an area at random, which is the reference area. Next, the geographically furthest area from the reference area is used as the initial seed for the first region. Then, neighboring contiguous unassigned areas are added to the seed in order of increasing distance from it. The adding process stops when the region satisfies a predefined condition, for example, a minimum population quota. When the first region is finished, the next region is designed by selecting as the initial seed the furthest unassigned area from the reference area. The strategy of having reference areas was implemented by Vickrey as a way to avoid the creation of

enclaves, which are the remaining unassigned areas that cannot be a region by themselves. Those enclaves must be assigned to an already created region.

Thoreson and Liittschwager (1967) expanded Vickrey's method in several aspects. First, they proposed repeating the regionalization process several times with different reference areas in order to explore new solutions. And second, they introduce a modification of the algorithm to be used on a regular lattice.

Later, Gearhart and Liittschwager (1969) introduced other improvements to Thoreson and Liittschwager's algorithm by including new termination rules, new ways to prevent the formation of enclaves, different alternatives to seed the regions, and a more formal, clearer presentation of the algorithm.

The seeded regions approach has also been studied by Taylor (1973), Openshaw (1977a, 1977b), and Rossiter and Johnston (1981). A common factor between these contributions is that they select the set of seeds, called "core areas," at the beginning of the algorithm. Then, neighboring areas are added to the cores. The process stops when all the areas are assigned. Rossiter and Johnston's method does not consider the next seed until a region has been finished, whereas Taylor and Openshaw's methods allow the regions to grow simultaneously.

Note that the algorithms derived from Vickrey's contribution assume that the number of regions is known. So, special attention must be paid to the criterion used for stopping the addition of areas to a region. On one hand, if the minimum population per region is set very low, then it is likely that the desired number of regions is reached very fast, with many areas waiting for assignment (enclaves). On the other hand, if the minimum population per regions is set very high, then the algorithm will not be able to create as many regions as wanted.

3.2.3. Modification of an initial feasible solution

The methods covered in this section require as an input an initial feasible solution which is then iteratively modified while searching for improvements in the aggregation criterion. A common factor among these methods is that each iteration must be feasible in terms of the spatial contiguity constraint.

This type of regionalization method was proposed by Nagel (1965) for redistricting problems. The process starts from an initial feasible solution that is modified by moving areas from their current region to another neighboring region. Several conditions must be satisfied before allowing

a move: first, the donor region (the region currently containing the area being moved) must have more than one area. Second, the donor region cannot lose its spatial contiguity after removing the area.

Nagel allows two types of moves. The basic one is to move one area at the time. The second type of move involves trading areas, on a one-for-one basis, between two neighboring regions.[17] These moves are allowed only if there is an improvement in the aggregation criterion. The process stops when no improving moves can be performed.

The contiguity constraint is tested by selecting one area within the region. Then neighboring areas belonging to the same region are added until no more areas can be added. If all the areas belonging to the region can be added, then the region is spatially contiguous.

The algorithm called *Automatic Zoning Procedure* (AZP) proposed by Openshaw (1977a) uses the same strategy, but it does not include trading moves. Although the methodological approach was not innovative with respect to earlier contributions, the main contribution made by Openshaw was to use these regionalization models to explore the effects of the Modifiable Areal Unit Problem (MAUP).

Up to this point all the algorithms were based on a hill-climbing procedure where only improving moves were allowed. Sammons (1978) highlighted the need to allow these procedures to perform non-improving moves in order to be able to explore a wider range of feasible solutions and escape from local optimal solutions, which are very likely to occur when these approaches are used. Sammons also introduced a very interesting additional way to explore more solutions based on merging/splitting regions: the number of merges must be equal to the number of splits to maintain the number of desired regions. In the same book, Openshaw (1978) presents an extension of Openshaw (1977a) where merging/split moves are incorporated as well.

Ferligoj and Batagelj (1982) studied this methodology within the context of clustering with relational constraint. The authors see the trading not only as useful for exploring potential improvements on the aggregation criterion, but also as a way of preserving the number of areas within each region. They also propose several ways to generate the initial feasible solution.

Browdy (1990) made an important contribution to this type of model. He formulates a model for the redistricting problem where non-improving moves are possible by implementing a

[17] When trading two areas, each must touch the neighboring region in an area other than the area being traded. This move cannot break the contiguity constraint of either of the two regions involved in the trading.

Simulated Annealing (SA) structure within the searching process. SA allows non-improving moves with a probability that diminishes gradually over iteration time. In this algorithm, special attention should be given to the definition of the initial parameters related to the cooling schedule to ensure an appropriate trade-off between the execution time and a good solution.

Macmillan and Pierce (1994) also used SA for redistricting problems in an algorithm called ANNEAL. They proposed an alternative method for checking for spatial contiguity called switching points, which proved to be more efficient than their method based on the powers of the binary contiguity matrix presented in section 3.1. The switching point method does not have to iterate over all the areas assigned to the region; it only focuses on the area that is the candidate to be removed from the donor region and its first order neighbors, including those neighbors assigned to other regions.[18] Empirical evidence on the performance of the switching points method is provided by Macmillan (2001).

Openshaw and Rao (1995) proposed two alternative approaches for improving Openshaw's AZP model. In one of the approaches, AZP-SA, non-improving moves of areas are allowed by using a SA scheme. The other approach, AZP-tabu, uses a well-known heuristic procedure called Tabu Search Algorithm,[19] which also attempts to escape from local optimal solutions by allowing non-improving moves. In AZP-tabu, once an area is moved from one region to another, its reverse move is prohibited for R subsequent iterations. R is then the length of the tabu list and its value is a key parameter in the algorithm. The authors also provide a version of AZP-tabu where the value of R is dynamically adjusted during the searching process.[20]

Horn (1995) also took up Sammons' idea of implementing different types of area moves and allowing for temporary changes in the number of regions. Horn modifies an initial feasible solution with three types of moves: Zone-at-march, moving an area from its region to a neighboring region; Territory-at-march, merging two regions; and Territory injection, selecting an area to create a new region. Here the number of regions can deviate from a predefined value in order to explore new solutions. The allowed maximum deviation from the desired number of regions is gradually reduced to zero to ensure a feasible solution. Finally, Horn stresses the

[18] The authors enumerate the special cases where this methodology may fail.
[19] For more information on the *Tabu Search Algorithm*, see Glover (1977, 1989, 1990).
[20] The possibility of dynamically adjusting the value of R is known as the Reactive Tabu Search (Battiti and Tecchiolli, 1994).

importance of implementing different types of moves and encourages the exploration of moves that involve more than one area at the time.

Duque and Church (2004) introduced the algorithm called Automatic Regionalization with Initial Seed Location (ARISeL) where special attention is paid to the design of a good initial feasible solution before performing a local search by moving areas between regions. The algorithm has two stages. The first stage uses a seeded regions strategy to generate an initial feasible solution. Information on how the aggregation criterion changes through the assignation process is used to make changes in the initial set of seeds. This first stage generates a set of feasible solutions from where the best solution is chosen and then improved in a second stage by applying a local search process based on a tabu search algorithm. Using a good feasible solution as an input to the second stage will reduce the possibility of getting trapped by a local optimal solution and also the number of moves performed during the second stage.[21]

3.2.4. Graph theory-based models

Maravalle and Simeone (1995) formulated a heuristic model for regionalization based on graph theory. The heuristic called MIDAS-*Méthode Itérative d'Agrégation Spatiale* represents the areas and their spatial contiguity relationships as a connected graph (G) where the link between nodes/areas i and j means that those two areas share a border. The problem is formulated as follows: *"Given a connected graph G, in which a vector of characteristics is associated with each vertex, find a minimum inertia partition of the vertex-set of G into a prescribed number of connected clusters"* (Maravalle and Simeone, 1995, p. 625).

The main steps of MIDAS are: first, generation of a spanning tree T of G[22], second, generation of an initial partition of T by deleting p-1 links where p is the number of regions to be designed, and third, the set of p-1 deleted links is iteratively modified by replacing one deleted link by another non-deleted link in T. This replacement, which is a neighboring feasible solution of T, is allowed if the new solution leads to an improvement in the aggregation criteria.

Since, for a given T, the number of neighboring solutions is very restricted, the authors introduce a fourth step called "tree-modification" which generates a different spanning tree $\overline{T}$ of G that results from replacing links in T by other unselected links available on G. This new $\overline{T}$

[21] The number of moves is reduced on average by 36.4% with respect to the total number of moves required by AZP-tabu.

[22] The authors applied Kruskal's algorithm in this step (Kruskal, 1956).

allows evaluation of new neighboring solutions that are not possible to evaluate from *T*. This process is repeated until there are no improving moves to perform.

The main advantage of this approach is the fact that spatial contiguity constraint is maintained through the iterations without having to test for it at each move, since deleting any combination *p*-1 links from any spanning tree *T* of *G* always leads to a feasible solution.

3.3. Mixed heuristic models

Under this category we classified the heuristics that combine heuristic and exact optimization models to solve regionalization problems. The aim of these models is to merge in one model the computational power of heuristic approaches with the mathematical accuracy of exact models. This combination has been made in two different ways: first, the exact model can be applied in a set of neighboring regions to redraw their borders, achieving improvements in the aggregation criteria; second, information obtained from multiple solutions generated by any regionalization heuristic is used to reduce the number of variables and constraints in an exact model.

The first method was proposed by Duque (2004). The algorithm called Regionalization Algorithm with Selective Search - RASS has as a main assumption that the design of contiguous and homogeneous regions is relevant only if there exist both disparities between the areas, which justify the design of more than one region, and some evidence of spatial dependence, which justifies the requirement of spatial contiguity. If these two properties are present in the data set, then the information available on the relationships between areas can be crucial in directing the search process in a more selective and less random fashion. The algorithm starts by selecting a subset of *m* neighboring regions. The areas belonging to those regions are passed to an optimization model[23] to re-aggregate them into *m* regions. Next, taking into account information on the relationships between the areas, the algorithm decides which region should leave the set of neighboring regions and which region should be added to the set in order to run the optimization model again. Thus, the set of regions passed to the optimization model keeps changing throughout the iteration process until a convergence criterion is satisfied.

The aim of RASS is to take advantage of the optimization model by applying it to a set of regions instead of trying to solve the whole problem at once. The local improvements achieved with the optimization model may be more difficult to obtain with an area-swapping scheme.

[23] Proposed by Duque (2004) and presented in section 3.1.

Duque and Church (2004) formulated an alternative way of taking advantage of the optimization model for regionalization problems. The authors propose the adaptation of a well-known heuristic model known as Heuristic Concentration – HC (Rosing and ReVelle, 1997). The central idea of this model is to use information from multiple solutions, obtained by any regionalization heuristic, in order to reduce the number of variables and constraints in the optimization model. The solution obtained from the reduced optimization model will be at least as good as the best solution obtained from the heuristic. The algorithm has two stages. The first stage runs a regionalization heuristic q times. The best m solutions are taken from the q available solutions. The second stage uses those m solutions to reduce an early version of the Order^{RM} model presented in section 3.1. Thus, if a set of areas is assigned to the same region in all the m solutions, those areas will be forced to be together in the optimization model. This information can also be used to reduce the index related to the maximum contiguity order (o). Empirical evidence shows that this procedure can reduce, on average, 83.8% of the constraints and 84.2% of the variables from the fully specified optimization model.

Middleton (2006) proposed a similar approach. He formulates a heuristic called Heuristic Distillation – HD to reduce the Flow^{RM} model proposed by Duque et al. (2006). The algorithm uses information from multiple solutions obtained from any heuristic model. Next, information from the best m solutions is used to reduce the Flow^{RM} model. Information on neighboring areas that are never assigned to the same solution, and on areas that are always assigned to the same region, is used to formulate a reduced version of the optimization model. This process can reduce, on average, 79.53% of the constraints and 56.51% the variables from the fully specified Flow^{RM} optimization model.

4. Conclusions

In this paper we review more than four decades of contributions to regionalization methods, and we have also propose a taxonomy of the different regionalization methods considered in the literature based on the strategy used to satisfy the contiguity constraint. Table 2 shows the main topics considered in the reviewed studies (column 2) following our proposed taxonomy (column 1). Taking this information into account, our aim in this study was to provide a big picture of the alternative methods that can be applied by researchers who see official aggregations as a source

of error when analyzing spatial data. This paper may also be a good starting point for researchers interested in developing new regionalization techniques.

TABLE 2 ABOUT HERE

A first conclusion from this review is that there is no ultimate regionalization technique. Each technique has characteristics that can be desirable in some applications but undesirable in other ones. In table 3 we provide an evaluation of those methods in terms of a set of characteristics. This table can be used as a guide for selecting a suitable regionalization method for a specific analysis.

TABLE 3 ABOUT HERE

Another conclusion is that, although there is a high degree of subjectivity in the criteria these methods use, the application of structured and replicable regionalization methods may reduce the criticism that the use of analytical regions usually receives.

Based on this review we envision various research lines that could contribute greatly to this field. First, elements from different regionalization strategies could be combined, leading to more powerful algorithms. Second, alternative ways of improving the computational tractability of exact optimization models for regionalization problems are needed. Third, less randomized aggregation processes could be designed by taking advantage of the available data (not only from classification variables but also from the spatial configuration of areas). Fifth, those models based on swapping or trading moves of areas could be extended towards the design of more aggressive moves that allow improvements in the local search process. And sixth, research addressed to assessing the real benefits of using analytical regions is needed.

References

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall.

Altman, M. (1998). *Districting Principles and Democratic Representation*. PhD thesis, California Institute of Technology.

Amrhein, C. G. and Flowerdew, R. (1992). The effect of data aggregation on a Poisson regression model of Canadian migration. *Environment and Planning A*, 24:1381–1391.

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

Arbia, G. (1989). *Spatial data configuration in statistical analysis of regional economic and related problems*. Dordrecht: Kluwer.

Bacao, F., Lobo, V., and Painho, M. (2005). Applying genetic algorithms to zone design. *Soft Computing*, 9(5):341–348.

Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W. P., and Vance, P. H. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, 46(3):316–329.

Battiti, R. and Tecchiolli, G. (1994). The reactive tabu search. *ORSA Journal on Computing*, 6(2):126–140.

Bixby, R. E. (2002). Solving real-world linear programs: A decade and more of progress. *Operations Research*, 50(1):3–15.

Bixby, R. E., Fenelon, M., Gu, Z., Rothberg, E., and Wunderling, R. (2000). MIP: theory and practice – closing the gap. http://citeseer.ist.psu.edu/bixby00mip.html (06/19/2006).

Browdy, M. H. (1990). Simulated annealing: an improved computer model for political redistricting. *Yale Law & Policy Review*, 8:163–179.

Bunge, W. (1966). Gerrymandering, Geography, and Grouping. *Geographical Review*, 56(2):256–263.

Byfuglien, J. and Nordgard, A. (1973). Region-building - a comparison of methods. *Norwegian Journal of Geography*, 27:127–151.

Cliff, A. D. and Haggett, P. (1970). On the efficiency of alternative aggregations in region-building problems. *Environment and Planning*, 2(3):285–294.

Cliff, A. D., Haggett, P., Ord, J. K., Bassett, K. A., and Davies, R. B. (1975). *Elements of spatial structure : a quantitative approach*. Cambridge [Eng.] ; New York : Cambridge University Press.

Cova, T. J. and Church, R. L. (2000). Contiguity constraints for single-region site search problems. *Geographical Analysis*, 32(4):306–329.

Current, J., Revelle, C., and Cohon, J. (1984). The shortest covering path problem - an application of locational constraints to network design. *Journal of Regional Science*, 24(2):161–183.

Desrochers, M. and Laporte, G. (1991). Improvements and extensions to the Miller-Tucker-Zemlin subtour elimination constraints. *Operations Research Letters*, 10(1):27–36.

Duque, J. C. (2004). *Design of homogeneous territorial units. A Methodological Proposal and Applications*. PhD thesis, University of Barcelona.

Duque, J. C. and Church, R. L. (2004). A new heuristic model for designing analytical regions. In *North American Meeting of the International Regional Science Association, Seattle*.

Duque, J. C., Church, R. L., and Middleton, R. S. (2006). Exact models for the regionalization problem. In *Western Regional Science Association annual meetings, Santa Fe.*

Eurostat (2006). Nomenclature of territorial units for statistics – NUTS. statistical regions of Europe. http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html (06/19/2006).

Everitt, B. S. (1993). *Cluster Analysis, Third Edition*. New York: Halsted Press.

Ferligoj, A. and Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika*, 47(4):413–426.

Fischer, M. M. (1980). Regional taxonomy - a comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics*, 10(4):503–537.

Fotheringham, A. S. and Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical-analysis. *Environment and Planning A*, 23(7):1025–1044.

Garfinkel, R. S. and Nemhauser, G. L. (1970). Optimal political districting by implicit enumeration techniques. *Management Science Series B-Application*, 16(8):B495–B508.

Gearhart, B. C. and Liittschwager, J. M. (1969). Legislative districting by computer. Behavioral *Science*, 14(5):404–417.

George, J. A., Lamar, B. W., and Wallace, C. A. (1997). Political district determination using large-scale network optimization. *Socio-Economic Planning Sciences*, 31(1):11–28.

Gillman, R. (2002). Geometry and gerrymandering. *Math Horizons*, 10:10–12.

Glover, F. (1977). Heuristic for integer programming using surrogate constraints. *Decision Science*, 8:156–166.

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(1):17–29.

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

Helbig, R. E., Orr, P. K., and Roediger, R. R. (1972). Political redistricting by computer. *Communications of the ACM*, 15(8):735–741.

Hess, S. W. and Samuels, S. A. (1971). Experiences with a sales districting model - criteria and implementation. *Management Science Series B-Application*, 18(4):P41–P54.

Hess, S. W., Weaver, J. B., Siegfeld, H. J., Whelan, J. N., and Zitlau, P. A. (1965). Nonpartisan political redistricting by computer. *Operations Research*, 13(6):998–1006.

Horn, M. E. T. (1995). Solution techniques for large regional partitioning problems. *Geographical Analysis*, 27(3):230–248.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Johnston, R. J. (1968). Choice in classification: The subjectivity of objective methods. *Annals of the AAG*, 58:575–589.

Kaiser, H. F. (1966). An objective method for establishing legislative districts. *Midwest Journal of Political Science*, 10(2):200–213.

Keane, M. (1975). The size of region-building problem. *Environment and Planning A*, 7(5):575–577.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kruskal, J. B. (1956). On the shortest spanning subtree and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50.

Lankford, P. M. (1969). Regionalization - theory and alternative algorithms. *Geographical Analysis*, 1(2):196–212.

Macmillan, W. (2001). Redistricting in a GIS environment: An optimization algorithm using switching points. *Journal of Geographical Systems*, 3:167–180.

Macmillan, W. and Pierce, T. (1994). *Spatial Analysis and GIS*, chapter Optimization modelling in a GIS framework: the problem of political redistricting, pages 221–246. Taylor & Francis.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.

Maravalle, M. and Simeone, B. (1995). A spanning tree heuristic for regional clustering. *Communications in Statistics-Theory and Methods*, 24(3):625–639.

Margules, C. R., Faith, D. P., and Belbin, L. (1985). An adjacency constraint in agglomerative hierarchical classifications of geographic data. *Environment and Planning A*, 17(3):397–412.
Martin, D., Nolan, A., and Tranmer, M. (2001). The application of zone-design methodology in the 2001 UK census. *Environment and Planning A*, 33(11):1949–1962.
Mehrotra, A., Johnson, E. L., and Nemhauser, G. L. (1998). An optimization based heuristic for political districting. *Management Science*, 44(8):1100–1114.
Middleton, R. S. (2006). *Geographical Distillation: Application of the p-Median, Traveling Salesman, and Regionalization Problems*. PhD thesis, University of California at Santa Barbara.
Miller, C. E., Tucker, A. W., and Zemlin, R. A. (1960). Integer programming formulation of traveling salesman problems. *Journal of the ACM*, 7(4):326–329.
Mills, G. (1967). The determination of local government electoral boundaries. *Operational Research Quarterly*, 18(3):243–255.
Murtagh, F. (1985). A survey of algorithms for contiguity-constrained clustering and related problems. *Computer Journal*, 28(1):82–88.
Nagel, S. S. (1965). Simplified bipartisan computer redistricting. *Stanford Law Review*, 17(5):863–899.
Openshaw, S. (1973). A regionalisation program for large data sets. *Computer Applications*, 3-4:136–147.
Openshaw, S. (1977a). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. *Transactions of the Institute of British Geographers*, 2(4):459–472.
Openshaw, S. (1977b). Optimal zoning systems for spatial interaction models. *Environment and Planning A*, 9(2):169–184.
Openshaw, S. (1978). *Spatial Representation and Spatial Interaction*, chapter An optimal zoning approach to the study of spatially aggregated data, pages 95–113. Leiden ; Boston : M. Nijhoff Social Sciences Division.
Openshaw, S. (1984). *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography, 38 (GeoBooks, Norwich).
Openshaw, S., Alvanides, S., and Whalley, S. (1998). Some further experiments with designing output areas for the 2001 UK census. School of Geography, University of Leeds. www.geog.leeds.ac.uk/papers/98-9/ (06/19/2006).
Openshaw, S. and Rao, L. (1995). Algorithms for reengineering 1991 census geography. *Environment and Planning A*, 27(3):425–446.
Openshaw, S. and Taylor, P. J. (1981). *Quantitative Geography*, chapter The modifiable areal unit problem, pages 60–70. London: Routledge.
Openshaw, S. and Wymer, C. (1995). *Census Users Handbook*, chapter Classifying and regionalizing census data, pages 239–270. Cambridge, UK, GeoInformation International.
Paelinck, J. H. P. (2000). On aggregation in spatial econometric modeling. *Journal of Geographical Systems*, 2:157–165.
Paelinck, J. H. P. and Klaassen, L. H. (1979). *Spatial Econometric*. Saxon House, England.
Perruchet, C. (1983). Constrained agglomerative hierarchical-classification. *Pattern Recognition*, 16(2):213–217.
Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357.

Rosing, K. E. and ReVelle, C. S. (1997). Heuristic concentration: Two stage solution construction. *European Journal of Operational Research*, 97(1):75–86.

Rossiter, D. J. and Johnston, R. J. (1981). Program GROUP - the identification of all possible solutions to a constituency-delimitation problem. *Environment and Planning A*, 13(2):231–238.

Sammons, R. (1978). *Spatial Representation and Spatial Interaction*, chapter A simplistic approach to the redistricting problem, pages 71–94. Leiden ; Boston : M. Nijhoff Social Sciences Division.

Segal, M. and Weinberger, D. B. (1977). Turfing. *Operations Research*, 25(3):367–386.

Shirabe, T. (2005). A model of contiguity for spatial unit allocation. *Geographical Analysis*, 37(1):2–16.

Spence, N. A. (1968). A multifactor uniform regionalization of British counties on basis of employment data for 1961. *Regional Studies*, 2(1):87–104.

Taylor, P. J. (1973). Some implications of spatial organization of elections. *Transactions of the Institute of British Geographers*, (60):121–136.

Thoreson, J. D. and Liittschwager, J. M. (1967). Legislative districting by computer simulation. *Behavioral Science*, 12(3):237–247.

Tobler, W. R. (1970). Computer movie simulating urban growth in detroit region. *Economic Geography*, 46(2):234–240.

Vickrey, W. (1961). On the prevention of gerrymandering. *Political Science Quarterly*, 76(1):105–110.

Weaver, J. B. and Hess, S. W. (1963). A procedure for nonpartisan districting - development of computer techniques. *Yale Law Journal*, 73(2):288–308.

Webster, R. and Burrough, P. A. (1972). Computer-based soil mapping of small areas from sample data .1. multivariate classification and ordination. *Journal of Soil Science*, 23(2):210–234.

Wise, S., Haining, R., and Ma, J. (2001). Providing spatial statistical data analysis functionality for the GIS user: the SAGE project. *International Journal of Geographical Information Science*, 15(3):239–254.

Wise, S. M., Haining, R. P., and Ma, J. (1997). *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling, and computational intelligence*, chapter Regionalisation tools for exploratory spatial analysis of health data, pages 83–100. Berlin [etc.]: Springer.

Zoltners, A. A. (1979). *Sales Management: New Developments from Behavioral and Decision Model Research*, chapter A Unified Approach to Sales Territory Alignment, pages 360–376. Marketing Science Institute.

Zoltners, A. A. and Sinha, P. (1983). Sales territory alignment - a review and model. *Management Science*, 29(11):1237–1256.

Tables and figures

Figure 1. Taxonomy of methods for solving regionalization problems

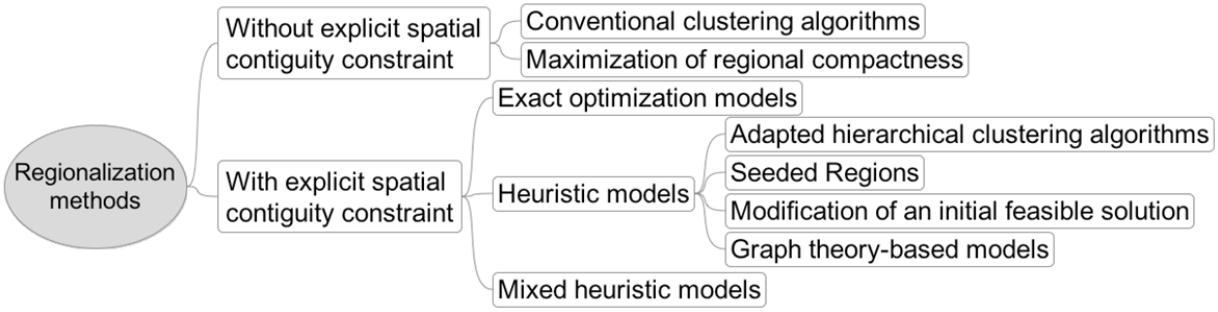

Table 1. Computational results for 14 counties in Vermont

	3 regions			9 regions		
	$Tour^{RM}$	$Order^{RM}$	$Flow^{RM}$	$Tour^{RM}$	$Order^{RM}$	$Flow^{RM}$
Rows-cols	1215-159	585 - 427	698 - 337	1215 - 159	1150 - 525	1863 - 767
OF	14.261	14.261	14.261	1.733	1.733	1.733
Best-Time (sec)*	29.16	47.04	120.8	0.37	9.56	79.29
Best-Gap*	60.63%	97.30%	24.79%	1.96%	100.00%	100.00%
Tot-time (sec)	16180.39	10800	147.55	0.37	10800	10800
Tot-gap	0.01%	37.15%	0.01%	0.01%	73.94%	100.00%

* Best-Time and Best-Gap: time and gap reported when the "Best-Integer" is equal to the optimal solution.

Table 2. Overview of the regionalization methods literature

Method	Topics addressed	Study
1. Regionalization via conventional clustering algorithms	Regionalization in two stages	Openshaw (1973) and Openshaw and Wymer (1995)
	Sensitivity of regionalization results to different clustering algorithms	Johnston (1968), Lankford (1969) and Fischer (1980)
	Simulated annealing variant of the k-means algorithm	Openshaw and Wymer (1995)
2. Regionalization via maximization of regional compactness	Consideration of natural boundaries	Mills (1967), Segal and Weinberger (1977) and George et al. (1997)
	Consideration of pre-existing aggregations	George et al. (1997)
	Fractional assignments	Helbig et al. (1972)
	Integer programming formulation	Hess et al. (1965) and Helbig et al. (1972)
	Quantitative measure of compactness	Weaver and Hess (1963)
	Types of moves of areas between regions	Kaiser (1966)
	Use of genetic algorithms	Bacao et al. (2005)
	Weighted combination of multiple aggregation criteria	Kaiser (1966), Cliff et al. (1975), Zoltners (1979) and Wise et al. (1997)
3. Exact optimization models	Alternative way to define adjacency between areas	Zoltners and Sinha (1983)
	Alternative ways to satisfy contiguity constraint	Duque et al. (2006)
	Complexity of the regionalization problem	Cliff and Haggett (1970), Cliff et al. (1975), Keane (1975) and Altman (1998)
	Contiguity constraint based on adjacency level	Zoltners and Sinha (1983)
	Contiguity constraint based on subtour breaking constraints	Duque (2004)
	Contiguity constraint based on the power of the contiguity matrix	Macmillan and Pierce (1994)
	Solution obtained by the union of predesigned feasible regions	Garfinkel and Nemhauser (1970) and Mehrotra et al. (1998)

(continued)

Table 2. (continued)

Method	Topics addressed	Study
4. Adapted hierarchical clustering algorithms	Comparison of hierarchical methods	Spence (1968), Webster and Burrough (1972), Byfuglien and Nordgard (1973) and Margules et al. (1985)
	Hierarchical methods and the identification of spatial patterns	Byfuglien and Nordgard (1973)

	Monotonic and nonmonotonic hierarchical clustering strategies with contiguity constraint	Ferligoj and Batagelj (1982)
	Reduction of similarity matrix	Openshaw (1973)
	Spatial contiguity based on contiguity threshold	Perruchet (1983)
	Sub-optimality of hierarchical algorithms	Bunge (1966)
5. Seeded Regions	Multi-start algorithm	Thoreson and Liittschwager (1967)
	Use of irregular and regular lattices	Thoreson and Liittschwager (1967)
	Ways to grow the regions	Vickrey (1961), Taylor (1973), Openshaw (1977a), Openshaw (1977b) and Rossiter and Johnston (1981)
	Ways to prevent enclaves	Gearhart and Liittschwager (1969)
	Ways to seed the regions	Gearhart and Liittschwager (1969)
6. Modification of an initial feasible solution	Types of moves of areas between regions	Nagel (1965), Openshaw (1977a), Openshaw (1978), Sammons (1978), Ferligoj and Batagelj (1982), Browdy (1990) and Horn (1995)
	Use of simulated annealing	Browdy (1990), Macmillan and Pierce (1994), Macmillan (2001) and Openshaw and Rao (1995)
	Use of tabu search	Openshaw and Rao (1995) and Duque and Church (2004)
	Ways to check for spatial contiguity (switching points)	Macmillan and Pierce (1994) and Macmillan (2001)
7. Graph theory-based models	Use of spanning tree	Maravalle and Simeone (1995)
8. Mixed heuristic models	Break the regionalization problem into sub-problems	Duque (2004)
	Use of heuristic concentration	Duque and Church (2004)
	Use of heuristic distillation	Middleton (2006)

Table 3. Regionalization methods and their main characteristics

Characteristics \ Regionalization method	Without an explicit spatial contiguity constraint		With an explicit spatial contiguity constraint					
	Conventional clustering algorithms	Maximization of regional compactness	Exact models	Adapted hierarchical clustering algorithms	Seeded regions	Modification of an initial feasible solution	Graph theory-based models	Mixed heuristics
Number of regions is required	✓	✓	✓	✓	✓	✓	✓	✓
Information about all pairwise relationships can be considered	✓	✓	✓	✓	✓	✓	✓	✓
Relationships between areas do not have to be geographically-dependent	✓	×	✓	✓	✓	✓	✓	✓
Neighboring relationships must be provided	×	×	✓	✓	✓	✓	✓	✓
Guaranteed spatial contiguity	×	×	✓	✓	✓	✓	✓	✓
Regional shape is not constrained	✓	×	✓	✓	✓	✓	✓	✓
Optimal solution is guaranteed	×	×	✓*	×	×	×	×	×
May be used to solve large regionalization problems**	✓	✓	×	✓	✓	✓	✓	×
May be easily adapted to any aggregation criterion	×	×	×	×	✓	✓	✓	×

* Only for Tour^{RM}, Order^{RM}, and Flow^{RM} models.

** 67 areas into 15 regions has been the largest problem solved by applying Heuristic Distillation to reduce the fully specified Flow^{RM} model (Middleton 2006). The reduced Flow^{RM} was solved with CPLEX 9.0 on a Sun Blade 2500. 500 MHz CPU, 2Gb RAM.