International Journal of Latest Research in Science and Technology
Volume 2, Issue 6: Page No.58-69 ,November-December 2013
https://www.mnkpublication.com/journal/ijlrst/index.php

ISSN (Online):2278-5299

GEOGRAPHIC CLUSTERING OPTIMIZATION WITH VARIABLE NEIGHBORHOOD SEARCH: A MULTIOBJECTIVE APPROACH

[1]María B. Bernábe,[2]Elías Olivares,[3] María A. Osorio, [4] Rogelio González, [5]Abraham Sánchez
[1,4,5]Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, México
[3]Benemérita Universidad Autónoma de Puebla, Facultad de Ingeniería Química, Puebla, México
[2]Universidad Popular Autónoma del Estado de Puebla,Centro Interdisciplinario de Posgrados, Puebla, México

Abstract*-Clustering is one of the most successful techniques for data mining, statistics, neuronal network, territorial design and others. In this kind of grouping, the parameters are usually optimized by means of a single objective.In particular, the partitioning is a clustering problem in the combinatorial optimization area and it has been well discussed and analyzed. However, real applications are far to be solved without the application of Multiobjective approaches. In this research paper we present a bi-objective partitioning proposal to solve the problem that involves census-based variables and geographical data for a territorial design problem.This is known to be a high complex computational problem and we it named Multiobjetive Clustering (MC).Two quality measures for clustering are chosen, which are simultaneously optimized in the partitioning process using Variable Neighborhood Search (VNS) for the optimization phase. The first quality measure obeys a geometrical concept of distances, whereas the second measure focuses in the calculus of the balance for a descriptive variable. In the multiobjetive clustering algorithm proposed(classification by partitioning), ithighlights a clear advantage with respect to the classical clustering algorithms such asK-meansand K-medoids which is the addition of another cost function which performs over variables vectors. The obtained results are shown in the Pareto frontier constructed with the approximate solutions generated by VNS, which are non-dominated and non-comparable with a similar mechanism on which the minimals of a Hasse Diagram and the Maxima Set are reached.

Keywords - Multi-objective optimization; Continuous Linear Time-Cost Trade-off; Bounded Objective Function Method; Construction Project; membership function

I. INTRODUCTION

Territorial Design (TD) occurs when small areas or geographical units must be bringing together into zones that result acceptable according to some requirements imposed by the problem in question. Geometrical compactness is a particular property which is often traduced as the clustering process in which a cost function is optimized. The current proposals of single-objective clustering in TD are diverse, from hierarchical algorithms to automatic classification (partitioning) [1, 2, 3]. The k-means clustering method is highly popular due to its simplicity for finding a partitioning of a data set into k disjoint groups. Here, a cost function is optimized by using a distance notion in the plane for data to be clustered. This restriction makes this problem to be located in the NP-complete category problems. Several clustering algorithms have been proposed in the basis of the k-means clustering method, with a common trade among them: the implicit or explicit optimization of a measure for the partitioning in which each group is represented by a centroid that attracts the other objects according to the given measure. However, due to the high computational cost of this problem, and for optimization purposes, the problem has been tackled by using different heuristics methods (see Bação et al. (2005)). In general terms, without specific constraints, usually grouping, clustering and classification are used similar mode. In fact, classification by partitioning is also known as automatic classification; however, a research line barely discussed in literature is the development of clustering methods for multiobjetive problems, which is focused in this research work. K-medoids is a combinatorial problem that has overcome some limitations of the k-means clustering method, however both algorithms keep the high computational complexity in the presence of other restrictions [4].

The partitioning is a method that belongs to clustering around medoids[5]. When the clustering process is performed around the medoids, the cost function is treatable as geometrical compactness which implicitly is satisfied by the properties of this type of partitioning. On the other hand, the major problems of TD demand other restrictions different than geometrical compactness such as homogeneity, contiguity and connectedness, which must be optimized simultaneously. In this point, for the particular problem of multi-objective clustering, we have focused our efforts in the simultaneous satisfaction of an additional restriction known as homogeneity for values or weights with descriptive variables, which is considered as another goal of interest along with compactness.

In this paper we introduce two quality measures expressed as cost functions that are simultaneously optimized. For the clustering process diverse basic aspects of k-medoids have been taken. The optimization process is carried out by using VNS due to its proved efficiency in difficult combinatorial optimization problems [6, 7]. The construction of the Pareto frontier is needed due to the fact that in this case we must

Publication History
Manuscript Received : 31 December 2013
Manuscript Accepted : 31 December 2013
Revision Received : 31 December 2013
Manuscript Published : 31 December 2013

optimize two cost functions. In order to obtain the approximate non-dominated solutions, the order theory was studied. We use the Hasse Diagrams with special attention in the minimal solutions and the maxima solution set, which both are non-dominated and non-comparable.

The rest of this paper is organized as follows. Section 2 presents the single-objective clustering algorithm. In Section 4 we describe the problem from a mathematical point of view, describing details of the cost functions for multi-objective clustering (MC) with VNS. In Section 3 the experimental results are discussed. Finally, in Section 6 the conclusions and future work are presented.

II. SINGLE-OBJECTIVE PARTITIONING

The classification techniques are of high usefulness when analyzing data. These techniques aim to bring together objects into classes or clusters, which internally are as homogeneous as possible, so that the objects that belong to the same cluster share a common trait. This level of similarity should be greater than the one that this objects have with respect to a different cluster. For practical purposes of the methodology proposed in this paper, it is necessary to have a measure D of dissimilarity among classification objects which may be expressed as a quality function. Even though many classification methods have been proposed throughout history, two broad categories can be distinguished: hierarchical methods and non-hierarchical methods or single link [8]. In this paper, we focused our attention in partitioning methods (non-hierarchical), in particular, based on medoids and iterative partitioning due to the strong influence in our MC algorithm. The main features of optimization methods is that they produce a single partition of k objects (a value specified beforehand) of non-overlapping clusters as a result of a maximization or minimization of one objective function [9]. Usually, these methods start with a initial k partition of the whole object set. Each partition has a centroid which is therefore used for attracting the rests of the objects. Each object is then re-localized to the closest centroid. The centroids are re-calculated and the process starts again until none change happens in all the clusters.

In general, the objects are represented by D descriptive attributes in the form of vector in the space R^D, and the similarity/dissimilarity measure used as distance, the cluster with similar objects are created. Due to in the process of constructing the cluster, none previous information about the structure of the cluster is given, this process is known as non-supervised classification. In clustering, variables of each object are used in order to measure the similarity among them.

In classification by partitioning we have $\Omega = \{x_1, \ldots, x_n\}$ as the finite set of n objects to classify and $k < n$ the desired number of classes into which the objects are classified. A partition $P = \{C_1, \ldots, C_k\}$ of into k classes, $C_1, \ldots, C_k$, is characterized by the following conditions:

1. $\Omega = \bigcup_{i=1}^{k} C_i$
2. $C_i \cap C_j = \emptyset$ for every $i \neq j$

The number k is the size of the partition. It's possible to eventually allow some of the C_i classes to be empty, in such a way that in reality the partitions $P = \{C_1, \ldots, C_k\}$ considered are partitions of into k classes or less. However, the optimal partitions according to the inertia criteria contain exactly k non-empty classes [10]. In general, the desire is to obtain classes as homogeneous as possible and such that they are sufficiently spaced.

Let P_k be the set of all the partitions of $P = (C_1, \ldots, C_k)$ of into k or less classes. Finding "good partitions" is desired, this is, those partitions that reflect the existent similarity between the objects $x_i \in \Omega$. Every object $x_i \in \Omega$ will be characterized by p different attributes or variables measured in a numerical scale, where each object x_i will be seen as a vector from the Euclidean space $\mathbb{R}^p$. In this representation space we have an Euclidean metric M (positive defined symmetrical matrix), that is used to define the internal product $\langle x_i | x_j \rangle_M = x_i^t M x_j$ between the objects and the norm $\| x \|_M^2 = x^t M x$. In the actual programming of the algorithms it's been assumed, without losing generality for convergence effects, that $M = Id$ (classic Euclidean metric). In effect, the general case M is decomposed as $U^t U$, and the transformation $z_i = U x_i$ takes us to the classic Euclidean metric, with the new data z_i.

The problem presented in this paper poses that given a set of n geographical objects $X = \{x_1, x_2, \cdots, x_n\}$, where $x_i \in R^D$, and k a positive number previously known. The problem of clustering consists of finding a partition $P = \{C_1, \cdots, C_k\}$ of X with C_j a cluster made up of similar objects, satisfying an objective function $f: \mathrm{P} \to R$, where is the collection of all the partitions of . In order to measure the similarity between two objects x_a and x_b, a distance function denoted by $d(x_a, x_b)$ is used. The Euclidean distance is the most useful similarity measure. Thus, the distance between two different elements $x_i = \{x_{i_1}, \cdots, x_{i_D}\}$ and $x_j = \{x_{j_1}, \cdots, x_{j_D}\}$ is given as shown in Eq. (1).

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{D} (x_{i_l} - x_{j_l})^2} \quad (1)$$

The objects belonging to one cluster are similar when the distance among them is minimal, this allows to formate the compactness objective function as shown in Eq.(2).

$$D(P) = \sum_{C \in P} \sum_{x_i \in C} d(x_i, x_j)^2, P \in \mathrm{P} \quad (2)$$

That is, it is required to minimize (2), where x_C is known as the representative element of the cluster C, which is the mean of the elements in the cluster C. This mean is calculated as shown in Eq. (3) and corresponds to the representative center of the cluster.

$$X_C = \frac{1}{|C|} \sum_{x_i \in C} x_i \quad (3)$$

The goodness of K-medoids based partitioning has developed diverse single-objective partitioning algorithms over geographical data [11]. In the same way, a MC multiobjective clustering algorithm presented in this paper, has inherit the characteristics of this type of single-objective partitioning. On the other hand, when the equation 2 has been minimized, is possible to understand that the compactness measure is implicitly satisfied. However, in an effort to express the geometrical compactness, we considered the classic partitioning definition adapting it to our problem [11]. In section 4.1 its description can be seen.

A. K-MedoidsAlgorithm

The K-means algorithm is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data. How might the algorithm be modified to diminish such sensitivity? Instead of taking the mean value of the objects in a cluster as a reference point, a Medoid can be used, which is the most centrally located object in a cluster. Thus the partitioning method can still be performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point. This forms the basis of the K-Medoids method. The basic strategy of K-Medoids clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (the Medoids) for each cluster. Each remaining object is clustered with the Medoid to which it is the most similar. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects. A typical K-Medoids algorithm for partitioning based on Medoid or central objects is as shown in Algorithm1:

Input:K: The number of clusters
Input:D: A data set containing n objects
Output:A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.

Arbitrarily choose k objects in D as the initial representative objects
Repeat
Assign each remaining object to the cluster with the nearest medoid;
Randomly select a non medoid object O_{random};
Compute the total points S of swap point O_j withO_{random};
If $S < 0$
Swap O_j with O_{random} to form the new set of k medoid
End If
Until no change; [12].

Algorithm1.A typical K-Medoids algorithm

B. The Variable Neighborhood Search (VNS) for partitioning

The VNS metaheuristic, proposed by Hansen and Mladenovic (1996 and 2003) is based on the observation that local minima tend to cluster in one or more areas of the searching space. Therefore when a local optimum is found, one can get advantage of the information it contains. For example, the value of several variables may be equal or close to their values at the global optimum. Looking for better solutions, VNS starts exploring, first the nearby neighborhoods of its current solution, and gradually the more distant ones. There is a current solution S_a and a neighborhood of order k associated to each iteration of VNS. Two steps are executed in every iteration: first, the generation of a neighbor solution of S_a, named $S_pN_k(S_a)$, and second, the application of a local search procedure on S_p, that leads to a new solution Sol. If Sol improves the current solution S_a, then the searching procedure will restart now from sol using $k = 1$. Otherwise, k is incremented and the procedure is repeated from S_a. The algorithm stops after a certain number of times that the complete exploration sequence $N_1; N_2; \cdots; N_{kmax}$ is performed (see Algorithm 2).

```
/* N_k : k = 1, ... , k_max, neighborhood structures*/
/* S_a:current solution*/
/* S_p: neighbor solution ofS_a*/
/* Sol:local optima solution*/
BEGIN
repeat k = 1; until END;
repeat
    /* Generate neighbor S_pof the k-th neighborhood
    of S_a(S_pN_k(S_a))*/
    S_p = GetNeighbor(S_a; N_k);
    Sol = LocalSearch(S_p);-
    ifSol is better than S_athen
        S_a = Sol;
    else
        k = k + 1;
    end
untilk = k_max;
END
```

Algorithm 2.Procedure Variable Neighborhood Search [13].

When the VNS method is incorporated in a partitioning algorithm, it obtains good results with respect to compactness and quality of solutions with a reasonable time of computing [11].

III.MULTI-OBJECTIVE PRELIMINARS

The multiobjective formulation is informally an optimization problem that presents two or more objective functions. The main inconvenient in this kind of problems in relation to a single objective model resides in the subjectivity of the solution found. A multiobjective problem doesn't have a unique optimal solution, generates a set of solutions that can't be considered different between the objectives it optimizes. The set of optimal solutions is denominated Pareto Frontier (PF) where the frontier of solutions contains all the points that aren't overcome in all of the objectives by another solution. This concept is denominated dominance, such that the PF consists only of non-dominated solutions, then, a solution dominates another one if and only if, it is at least as good as the other in all of their objectives and is at least better in one of them [14].

The precision of these problems can be given can be achieved if the relationships between its characteristics, its restrictions and the objectives are identified, then it is possible to express it as a mathematical function. The improvement all together, means that all the functions must be optimized simultaneously, which leads to the following definition:

Definition 3.1A multiobjective problem (MOP) can be defined in the case of minimization (and analogously for the case of maximization) as follows: Minimize f(x)

such that $f: F \subseteq R^n \to R^q, q \geq 2$ with feasible region in $A = \{a \in F: g_i(a) \leq 0, i = 1, \cdots, m\} \neq \emptyset$

The set is known as feasible region and it is said that the problem is subject to the restrictions $g_i: R^n \to R$ that are any functions.

In multiobjective optimization a certain scheme has to be selected that defines the improvement of one solution over another, frequently known as domination scheme and its definition is mainly based in that the solution of a multiobjective problem it is not unique and therefore the decision maker must choose one among a range of possible solutions that cannot out best each other, that is, that they don't dominate each other. The clarity of this concept is given when we think that within the real numbers field the order is defined in a natural way. For R^n, is possible to extend the concept by means of the following definition.

Definition 3.2 Given x, y vectors in R^n with $x \underline{\prec} y$ if and only if $x_k \leq y_k$ for every $k \in \{1, \cdots, n\}$ and $x \prec y$ if and only if $x \underline{\prec} y$ with $x \neq y$, where is the usual order in .

A. Pareto Frontier

The dominance relationship is known as Pareto dominance defined as follows:

Definition 3.3 Given a multiobjective problem, minimize $f(x)$, where $f: F \subseteq R^n \to R^q, q \geq 2$ with $A \subseteq F$ the feasible region. We say that a vector $x^* \in A$ is non-dominated or an optimal Pareto if there is no vector $x \in A$ such that $x \prec x^*$.

Thus, the answer to the problem of finding the best solutions (the non-dominated solutions, however the domination is defined within the technique) in a multiobjective problem is what is known as the solution set of the problem and the set of values of the objective function with a domain restricted to the vectors of the solution set (that is, the non-dominated vectors) is what we know as Pareto Frontier. In this way, the concept of set of non-dominated vectors, logically leads to the concept of partially ordered set.

Definition 3.4 The set $E(A; f)$ of Pareto efficient solutions (also known as set of Pareto optima) is defined as follows:

$E(A, f) = \{a \in A: \exists b \in A \text{that satisfies } f(b) \prec f(a)\}$

That is, the set of all the non-dominated vectors under the Pareto scheme. Summing up, the set Pareto optimum is the solution space of the problem, and the Pareto Frontier is its image in relation to the function to optimize.

$$f: F \subseteq R^n \to R^q, q \geq 2$$

A concept closely related with the Pareto Frontier is the one of Pareto optimum. The Pareto optimum as well as the Pareto Frontier are the frame over which the multicriteria decision making is worked. The set of Pareto optimum for a given multiobjective problem, is a partially ordered set (poset) seen in a formal way. In the multiobjective problems, the minimal elements are searched for in the solution space seen as a poset with the relation $\preceq$ given in definition 3.2. At this point a brief parenthesis over the concepts of order has been relevant to achieve the PF that we incorporate to the process of MC with VNS:

Definition 3.5 Given a set and ( ) a partial order relationship over it, we call the couple $(A, \prec)$ a partially ordered set also referred as Poset.

Definition 3.6 Given $(A, \prec)$ a Poset, the subset $X \subseteq A$ it is said to be a total order or chain in relation to ( ), if and only if $x \prec y$ or $y \prec x$ is fulfilled for every $x, y \in X$. In this case it is said that $(X, \prec)$ is a totally ordered set.

From a partial order the domination relationship ( ) can be defined in the following way: $x \prec y \Leftrightarrow x \prec y \wedge x|negy$ When occurs that $x \prec y \wedge y \prec x$ it is said that they are not comparable, which is denoted by $x \parallel y$.

On the other hand is important as well citing the lemma of Zorn, also known as Kuratowski-Zorn [15], it is a proposition of the sets theory that states the following:

Any not empty partially ordered set in which any chain (totally ordered subset) has an upper bound, contains at least one maximal element.

A maximal element of a partially ordered set is an element that is not lower than any other element. The term minimal element is defined in the dual way.

Definition 3.7 Let $(P, \prec)$ be a partially ordered set; $m \in P$ is a maximal element of if the only $x \in P$ such that $m \prec x$is $x = m$. The definition of a minimal element is obtained by changing the symbol by .

At first sight it would seem that m should be a maximum element, which is not always true because the definition of a maximal element is something weaker. In fact, maximal elements can exist even if there is no maximum. The reason is that, in general, is only a partial order in ; if is a maximal and $p \in P$, there is the chance that neither $p \prec m$ nor $m \prec p$, therefore would not be a maximum. This allows, furthermore, that there might be more than one maximal element in a set.

However, if $m \in P$ is maximal and has a maximum, it will be fulfilled that $\max(P) \prec m$; by definition of maximum it must have $m \prec \max(P)$ and therefore $m = \max(P)$; in other words, a maximum, if exists, it is the only maximal as well.

It is not hard to see that if is a total order in , the notions of maximum and maximal, coincide. Let $m \in P$ be a maximal element, and $p \in P$ arbitrary; by the condition of total order, either $p \prec m$or $m \prec p$; in the second case it would be $p = m$ by the definition of maximal, by which $p \prec m$, and therefore, $m = \max(P)$.

Not always the maximal elements exist, not even in the case where is totally ordered [16].

Maximal and minimal elements in a Hasse diagram

Let $(A \prec)$ be a CPO. Let $m, n \in A$. Then

1. is a maximal element in if and only if $(\forall x)(n \prec x \Rightarrow n = x)$
2. is a minimal element in if and only if $(\forall x)(x \prec m \Rightarrow x = m)$

Intuitively, an element of a CPO $(A \prec)$ is maximal of if there is no element in that is strictly greater than .

Analogously, an element of is a minimal of if there is no element of that is strictly less than .

Variable Neighborhood Search in the multiobjective partitioning "implicitly identifies branches" of the Hasse diagram where each non-dominated point of each branch is the maxima set (that contains all of the maximals) and is the Pareto Frontier and is the set of non-comparable minimal points, that is, is the Maxima set (Minima its dual) [17].

B. Approach to the problem

The problem being treated here is about Multiobjective Partitioning in DT and the main conflict is to optimize simultaneously the functions of compactness for the geographical location and homogeneity for census variables. As a case of study, the data to group are Basic Geo-statistical Areas (AGEBs by its initials in Spanish) and answers to different population problems about concentration or distribution of a census value in a determined metropolitan zone are searched for. This distribution has a value associated that must be balanced for each of the groups of the whole territorial extension (homogeneity) and at the same time respect the property of physical proximity between the Agebs that form a group (compactness). The approximation of the cost function is done with VNS and it is combined with a method that supports on the order theory to find a set of non-dominated solutions and non-comparable through the minimal points of a Hasse diagram that form the Maxima set [18]. It has been proposed a variant to the Pareto order relationship (see definition 3.2, 3.3 and 3.4). This variant guarantees finding non-comparable and non-dominated solutions, with a simple strategy: evaluate that the solutions generated by VNS over compact and homogeneous partitioning, meet the Pareto order and at the same time that they are non-comparable.

Translating the Pareto order we have that:

Given a solution (a, b) the next solution (a', b') is accepted if: $(*)$

$$a' > a \wedge b' = b \vee b' > b \wedge a' = a \vee a' > a \wedge b' > b \vee a = a \wedge b = b'$$

The trivial implication of logically negating the expressions gives place to obtain the following: Let's consider inequality again

$$(a, b) < (a', b')$$

Then the negation of this relationship produces the following equivalences:

$$\begin{aligned} &\neg((a, b) < (a', b')) \equiv \neg[(a < a' \vee a = a') \wedge \\ &(b < b' \vee b = b')] \equiv \\ &\equiv \neg(a < a' \vee a = a') \vee \\ &\neg(b < b' \vee b = b') \equiv \\ &\equiv (a \geq a' \wedge a \neq a') \vee \\ &(b \geq b' \wedge b \neq b') \equiv \\ &\equiv (a > a' \vee b > b) \end{aligned} \quad (4)$$

In the same way we have

$$\neg((a', b') < (a, b)) \equiv (a' > a \vee b' > b)$$

Therefore, it is concluded that (a, b) y (a', b') are non-comparable if $(a > a' \vee b > b') \wedge (a' > a \vee b' > b)$. Under this non-comparable order properly adjusted with the Pareto Order all the pairs of minimal solutions are obtained satisfying in this way the conflict of the compromised solutions.

In general terms, our algorithm generates a history of all the minimal candidate solutions. This process is repeated until the VNS' parameters allow it, thus collecting the set of minimal solutions that from the Pareto Frontier. The contribution in this work is the improvement of the algorithm reported in [18]. The challenge has been obtaining the exact Pareto Frontier of the solutions filtered with Nodom[17]. On the other hand, the result we have presented in previous works consisted in revealing a list of minimals that from the PF. However, only one solution from these minimals, the best found, was explicit in the sense that it showed the clusters and the objects that belonged to it. The improvement of the exposed algorithm in this paper not only resides in showing the PF without additional solutions "even if they are close to the PF", the program creates a more complete output file, where each minimal solution has its cluster associated, compactness-homogeneity cost and files to create the maps. This result benefits considerably to the multicriteria decision maker.

IV. MULTI-OBJECTIVE VNS

MC with VNS consists in the clustering of geographical objects in which a bi-objective function simultaneously minimizes compactness for the geographical location, and homogeneity over the descriptive variables. In MC, the creation of groups/clusters considers geographical data in the aggregation, defined by a variable vector with population attributes. The clusters represented by two cost functions are subject to the satisfaction of the minimization process with VNS: a distance measure in the geographical space and another distance of balance for the population variables.

Optimization with VNS is reduced to search the minimum for each solution (partition generated), which is made up by a pair of values (compactness, homogeneity) denoted by (c_i, h_i), where $i \leq$ number_of_objects. By a process, analogous to the obtaining of Hasse Diagram minimals, the solutions (c_i, h_i) are paralleled evaluated with a special order relationship named non-comparable order, which is integrated into the Pareto Dominance (PD). This procedure obtains a set of non-comparable and non-dominated pairs (c_i, h_i), i.e., minimal solutions that form the Pareto frontier:

A. First objective: Geometrical compactness

Even though many authors in literature have dedicated efforts for describing compactness in a quantitative manner, this concept has not been defined in a precise manner. In [19] more than 20 different measures may be reviewed. Intuitively, when we talk about geometrical compactness in a territory, we think that each zone cluster resembles a convex geometrical figure such as a circle or a square. This leads to consider that the measures that have been proposed up to now are not totally convincing [20]. Some examples of the compactness measures proposed are shown in Equations (5), (6), (7) and (8).

$$\sum_j \sum_{i \in Z_j} d_{i,j} \quad (5)$$

where $d_{i,j}$ represents the Euclidean distance between the - th geographical unit and the center of the zone that contains it.

$$\sum_i \frac{pr_i^2}{a_i} \quad (6)$$

wherepr_i is the perimeter of the zone, and a_i is the area for the -th zone.

$$\frac{\sum_{j\in J} R_j(x) - R}{2R} \quad (7)$$

where$R_j(x)$ is the perimeter of the zone in the solution , and is the perimeter of the total territory.

$$\sum_{i\in J} \frac{1 - \frac{2\pi\sqrt{\frac{A_j(x)}{\pi}}}{R_j(x)}}{m} \quad (8)$$

where$R_j(x)$ is the perimeter of the zone in the solution , and $A_j(x)$ is the circle that has the same area than the zone .

Some of the characteristics that a good compactness measures should have are the following:

1. The measure should be applicable to all the geometrical forms: regulars and irregulars.
2. The measure must be independent of transformations such as scaling and orientation changing.
3. The measure must be dimensionless, and being defined preferably in a scale between 0 and 1, with 1 describing a highly compact region.
4. The measure does not have to be affected by one or two extreme points.
5. The measure must correspond with the "intuition".

For the problem addressed in this paper, we informally say that there exist compactness when various objects "closely meet" so that there are no gaps, or these gaps are so few, that it meets that the distance among the objects toward the centroid representing the group be minimum, achieving implicitly clusters without geometric deformations. Therefore, when the objects are geographical, it is necessary to define a distance measure between clusters. Let be $X = \{1,2,\cdots,n\}$ the set of objects to be classified, the task consists of splitting into clusters $X = \{C_1, C_2, \cdots, C_k\}$with $k < n$, so that:

- $\bigcup_{i=1}^{k} C_i = X$
- $C_i \bigcap C_j = \emptyset$with$i \neq j$
- $|C_i| \geq 1$with$i = 1,\cdots,k$

A cluster C_m with $|C_m| > 1$ is compact if for each object $t \in C_m$ is fulfills Eq.(9).

$$\mathrm{Min}_{i\in C_m, i\neq t} d(t,i) < Min_{j\in Z - C_m} d(t,j) \quad (9)$$

A cluster C_m with $|C_m| = 1$ is compact if its object $t \in C_m$ is fulfills Eq.(10).

$$\mathrm{Min}_{i\in Z-\{t\}} d(t,i) < Min_{l\in C_f, \forall f\neq m} d(t,l) \quad (10)$$

From this criteria and the formulation described in Eq.(2), let say $f_1(x)$, it we can state that this geometrical compactness function is correct for the problem in question.

B. Second objective: homogeneity

Different methods for calculate the population balance among the clusters has been proposed. Even if some proposals are similar there is no evidence about the advantages when a particular method is applied. Some useful formulas for the homogeneity calculation follow:

1. The simplest way for measuring population balance consists on summing the absolute values of the difference between the population of each zone and the average population by zone.

$$\sum |P_i - \overline{P}|$$

Where P_i is the population of the zone , and $\overline{P}$ is the average population by zone calculated as $\overline{P} = \sum_{k\in K} \frac{P_k}{n}$, with equal to the number of zones to create, is the set of all the geographical units, and P_k is the population of the -th geographical unit.

2. The population difference between the most populated zone and the less populated one.

$$\mathrm{MAX}^{P_i} - \mathrm{MIN}^{P_k}$$

In some situations, this difference is divided by the average population as follows:

$$\frac{\mathrm{MAX}^{P_i} - \mathrm{MIN}^{P_k}}{\overline{P}}$$

3. The division of the most populated zone between the less populated one.

$$\frac{\mathrm{MAX}^{P_i}}{\mathrm{MIN}^{P_k}}$$

4. The following method is given by the function

$$\frac{\sum_{j\in J} \max\{P_j - (1+\beta)\overline{P} - P_j, 0\}}{P}$$

where is the set of all the zones, P_j is the population in zone , $\overline{P}$ is the average population by zone, is the percentage of population standard deviation. In this way, it is intended that the population of each zone is between the interval $[(1-\beta)\overline{P}, (1+\beta)\overline{P}]$, with $0 \leq \beta \leq 1$. This function will take the value of zero if the population of each zone is in the interval $[(1-\beta)\overline{P}, (1+\beta)\overline{P}]$, otherwise, it will take a positive value equal to the sum of the standard deviations with respect to this bounds [21].

For the MC problem presented in this paper, the second objective consists of finding a balance of an interest variable (homogeneity variable) for the geographical objects described by quantified attributes. After pre-processing the data, the filtered variables that are involved in the MC process are obtained. These variables can be: a) All the variables without restrictions, b) all the bounded variables, c) some variables without restrictions and the remaining do not participate, d) some variables with restrictions and the remaining do not participate.

The process for selecting these variables is described through a participation matrix (see Definition 4.1).

Definition 4.1 Let be $\Omega' = \{OG_1, OG_2, \cdots, OG_n\}$ a set of objects, and $VC = \{X_1, X_2, \cdots, X_r\}$ a set of census variables that describe the objects. Each variable X_i is a function of the object set Ω' with positive and real values R^+. Given intervals $I_j = [\alpha_j, \beta_j]$, $j = 1,\cdots,r$, and the characteristic functions $\chi_{[\alpha_j,\beta_j]}: VC \to \{0,1\}$,

$$\chi_{[\alpha_j,\beta_j]}(X) = \begin{cases} 1, & \text{ifx} \in [\alpha_j, \beta_j] \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

The matrix of participation associated to the cluster of objects Ω' with variables VC and conditions l_j $(j = 1, \cdots, r)$ is defined as the matrix

$$M = (v_{i,j})_{n\times r}, \quad (12)$$

where $v_{I,j} = \chi_{[\alpha_j,\beta_j]}(X_j)X_j(OG_i)$.

Therefore, the matrix contains all the values of the variables participating in the respective objects. If $v_{i,j} = 0$, then we say that the variable X_j do not participate in the object OG_i. The following step is to homogenize the clusters. By analyzing different homogeneity proposal, we have noted that there is sensibility in the results, which implied to propose an ad-hoc homogeneity measure for the current problem as follows. The ideal average for the interest variable is obtained. Let assume that the variable of interest is X_j and that its ideal average is V_j, (this phenomenon occur when all the clusters have the same value). Of course, this results if uncommon, and, therefore, it leads to propose a more sensible measure that calculate the real average for each group $(\frac{1}{n}\sum_{i=1}^{n} v_{i,j})$ subtracting this value from the ideal average as shown in Eq.(13).

$$V_j - \frac{1}{n}\sum_{i=1}^{n} v_{i,j} = \frac{1}{n}\sum_{i=1}^{n} (V_j - v_{i,j}) \quad (13)$$

By minimizing this difference in absolute value, the cost for the homogeneity cost function is acquired. The compactness cost is given by Eq.(2), whereas the homogeneity cost is given by Eq.(13) and denoted by $f_2(x)$.

Let $f_1(x)$ be the compactness function that requires the object geographical location for obtaining the distance matrix with the Euclidean measure (Eq. 2), which is the input for the function $f_1(x)$. Let $f_2(x)$ be the homogeneity function with the participation matrix (obtained by Eq. 13) as input of the algorithm. This calculus needs the quantified variable set associated to each geographical object.

The MC algorithm with VNS adopts the following restrictions:

1. Each object must belong to a single cluster
2. The value of each parameter in a cluster is the value of the census variable
3. The clusters are disjoint
4. There are not empty clusters
5. The descriptive variables may or not be bounded
6. The cluster may contain all or a subset of variables
7. The objects assigned to each cluster must conform a compact cluster (compactness cost function)
8. The clusters must be balanced with respect to the goal of balance for some measurable characteristic (homogeneity cost function)

In order to implement the algorithm VNS for MC, the aspects of partitioning with descending VNS has been integrated. $y = f(x) = (f_1(x), f_2(x))$has been simultaneously optimized with the Pareto dominance [18]. The algorithm is shown in the following subsection.

When multiple objectives are required to be optimized, the evident difficult arises due to the search of efficient optimal and non-dominated partitions. This particular problem is hardly treated in the literature. The variable homogeneity is an additional restriction required together with compactness for a demographic problem of DT. In this paper, we deal with the problem of MC maintaining the properties of single-objective clustering and including the homogeneity as an extension of the clustering method. Once the partitions have been constructed under the minimization of distances, the homogeneity is calculated for the compact partition already created. In [22] we may read the following affirmation: "For multi-objective problems, the cost function to be optimized has the same domain for all the objectives". This declaration means that for our particular problem, the two objectives are optimized over the same partition.

In MC, the local optima are avoided by using VNS. In order to obtain the non-dominated solutions the properties of the maxima solution set has been exploited [17].

C. The multi-objective clustering with VNS algorithm

The following algorithm, is a procedure that has been taken from [18], where the first version of MC was exposed. In this work we have improved the approaching to the pareto frontier where the all the generated solutions are non-dominated and non-comparable. In the previous work a couple of non-dominated solutions went through. Precisely the instruction

CurrentCost ← getCompCost(CurrentSol, getHomCost(CurrentSol))has been adjusted with the Pareto Dominance order, and with a relationship of special order that is described in the following section.

```
/* Let be:          */
/* n The number of objects to classify          */
/* k The number of groups*/
/* Val_i  The value that has the AGEB i for the variable that the homogeneity will be kept for*/
/* U_g  The geographical unit*/
/* MaxV NS The number of times that the neighborhood structures will be run through*/
/*MaxBL The maximum number of iterations for local search*/
Begin
kNeighborhood← Generate a random number between 1-n ;
CurrentSol← Generate a random solution that is found in the neighborhood kNeighborhood ;
CurrentCost←getCompCost(CurrentSol),getHomCost(CurrentSol) ;
cont← 1 ;
whilecont<MaxV NS do
   kNeighborhood←1 ;
   whilekNeighborhood≤ n do
         SolCand←Generate a random solution that is found in the neighborhoodkNeighborhood ;
         SolCand←LocalSearch(SolCand);
         CostCand←getCompCost(SolCand),getHomCost(SolCand) ;
         if(costCand<CurrentCost) then
               CurrentSolSolCand ;
```

```
                CurrentCost-costCand ;
            else
            kNeighborhood←kNeighborhood + 1 ;
            end
    end
    cont←cont+1 ;
end
Return CurrentSol ;
```

Algorithm 3. The multi-objective clustering with VNS algorithm

```
NumLoops←0 ;
BetterSol←Sol ;
costSolImp←getCompCost(BetterSol),
getHomCost(BetterSol) ;
whileNumLoops<MaxBLdo
SolCand←Generate random neighbor solution of BetterSol ;
costCandComp←getCompCost(SolCand),
getHomCost(SolCand) ;
    if (costCandComp≤costSolImp ) then
    BetterSol←SolCand ;
    NumLoops←MaxBL + 1 ;
    else
    NumLoops←NumLoops + 1 ;
    end
end
Return BetterSol ;
```

Algorithm 4: The local search algorithm: Function LocalSearch(Sol)

```
/* Returns an integer value that indicates how good is the
solution Sol with respect to compactness (the smaller the
value the better the solution)          */
i←1 ;
cost←0 ;
while i ≤ n do
    ifUg_i is not a centroid then
      /* Distance between Sol_1and the object i*/
      dmin←dist(Sol_1;Ugi) ;
    j← 2 ;
    while j ≤ k do
    ifdist(Sol _j;Ug_i) <dminthendmin←dist(Sol _j;Ug_i) ;
    end
    j←j + 1 ;
            end
            cost←cost + dmin ;
    end
    i←i + 1 ;
end
getCompCost(Sol) ← cost ;
```

Algorithm 5: Function getCompCost(Sol)

```
/* Returns an integer value that indicates how good is the
solution Sol with respect to homogeneity (the smaller the
value the better the solution)*/
Total←0 ;
cost←0 ;
forі ←1 to ndo
    ng← Get the number of the group to which the AGEB i
    belongs to ;
    total←total +Val_i ;
    totalGroup_ng←Val_i ;
end
IdealAverage← total/k ;
forj←1 to k do
    cost←cost + |jtotalGroup_j–IdealAveragej |;
end
getHomCost(Sol)←cost ;
```

Algorithm 6: Function getHomCost(Sol)

V. MODELING

We have pointed out that we are interested in finding partitions of (AGEBs) that minimize the compactness and homogeneity, some small adaptations to the definitions 4.1, 3.2 and 3.3 are required. For this we consider the collection of all the partitions of :

$$P = \{P: P \text{ is a partition of } \Omega\}$$

Let $f: P \to R^2$ be the function such that $f(P) = (C(P), H(P))$ where y are the compactness and homogeneity functions respectively, both with domain in and values in . The function of compactness is given by:

$$C(P) = \sum_{C \in P} \sum_{i,j \in C} d(i,j) \qquad (14)$$

Analytically the function of homogeneity , the second objective, has been described in the equation 13 as

$$V_j - \frac{1}{n}\sum_{i=1}^{n} v_{i,j} = \frac{1}{n}\sum_{i=1}^{n} (V_j - v_{i,j}) \qquad (15)$$

In our case the definition (3.1) is reduced to the following multiobjective problem: Minimize f(P) such that $f: P \subset 2^{\Omega} \to R^2$, with feasible region in $P = \{P \in 2^{\Omega}: P \text{ is partition of } \Omega\}$, where 2^{Ω} is the power set of and $f(P) = (C(P), H(P))$

Given the previous multiobjective problem we can include a partial order $\preceq_P$ over the set of partitionsP in the following way: $P \preceq_P P'$ if and only if $f(P) \preceq f(P')$, where $\preceq$ is the order given in definition 3.2. Analogously to definition 3.3, we say that a partition $P^* \in P$ is non-dominated or a Pareto optimum if there is not a partition $P \in P$ such that $P \prec_P P^*$, where $\prec_P$ is the strict order induced by the partial order $\preceq_P$.

Then the set of Pareto optima $E(f, P)$ in our case is definedas: $E(f, P) = \{P \in P: \exists P' \in P \text{ that meets } P' \preceq_P P\}$. Observe that the set of the partitions P is generated from the finite set then the image (Pareto Frontier) of the objective function is finite, and thus the Pareto Frontier is a discrete set. The goodness of the mechanism of our work to search for solutions of better compromise resides in the way the grouping is solved: it returns a set of diverse partitions with the use of VNS. On the other hand, to find the set of efficient solutions and non-dominated, the solutions generated along the process are evaluated, checking that they are non-dominated and non-comparable.

Finally the optimization for the objective of compactness has been solved by a partitioning algorithm based in the method of -medoids.

Given a partition $P \in P$ for each $C \in P$ we choose in a random way a $c \in C$ and define the sum

$$S(P) = \sum_{C \in P} \sum_{i \in C} d(i, c) \qquad (16)$$

Then the number

$$\text{Min}\{S(P): P \in P\} \qquad (17)$$

minimizes the intra class distance between AGEBs. Having as restrictions:

- $C \neq \emptyset$ (the groups are not empty).
- $C \cap C' = \emptyset \text{for} C \neq C'$ (there are not AGEBs repeated in different groups).
- $\bigcup_{C \in P} C = P$ (the union of all the groups consists of all the AGEBs).

The random choice of the centroids $c_1, \cdots, c_k$ generates a partition $P = C_1, C_2, C_k$ where each c_i is a representative of the class C_i

This partition is built in the following way:

1. An element $i \in \Omega$ is chosen.
2. The $\min\{d(i, c_t): t = 1, \cdots, k\}$ is calculated.
3. is located in the class C_t, where C_t, is the centroid, where the $\min\{d(i, c_t): t = 1, \cdots, k\}$ is reached.

Then partitions are generated, as many as the random selections of centroids made. The number of random selections is the number of iterations and is denoted by . Because the number of partitions of can be very big [10], the formulas (14) and (17) are restricted to the subset P' of of all the partitions generated from the different selections of the groups of centroids. Observe that the cardinality of P' is the number of iterations . Depending on the type of problem is necessary to fix a number of groups that the geographical zone will be partitioned into, this is, each element of the set P' has the same size as . Therefore P' is a set of all the partitions of size formed by selections of groups of centroids. The second objective, consists in the Minimization of homogeneity for a census variable (Equation 18)

$$V_j - \frac{1}{n}\sum_{i=1}^{n} v_{i,j} = \frac{1}{n}\sum_{i=1}^{n} (V_j - v_{i,j}) \qquad (18)$$

The model in question is mixed integer and makes use of the binary variables for models of the kind. Considering the above, the general model is:

$$\text{Minimize} \quad y = f(x) = (f_1(x), f_2(x)) \qquad (19)$$

f_1: is the cost of minimizing the distance between AGEBs according to equation 14 that must be formulated as function, and

f_2: is the cost of homogenizing (minimizing the homogeneity) a census variable of the AGEBs. Considering $f_1 = y_1$and $f_2 = y_2$, this function can be expressed as $y = (y_1, y_2) \in Y \subset R^2$ is the objective vector.

In the next subsection the final algorithm for MC with VNS in pseudocode is shown, as can be observed, unlike the algorithm in [18], presented in Section 4, a refinement to obtain the minimals has been achieved. (See Function UpdateMaximals(Comp, Hom)).

Due to the fact that the literature about order theory underlines the term maximals, this term has been kept in the pseudo code. Of course, for the problem we have solved, by minimizing 2 functions, by duality it is assumed that we obtain minimals in PF.

A. VNS for multi-objective clustering

```
/* Let be:                                              */
/* n The number of objects to classify                  */
/* k The number of groups                               */
/* Vali  The value that has the AGEB i for the variable that
the homogeneity will be kept                            */
/* Ug  The geographical unit                            */
/* MaxV NS The number of times that the neighborhood
structures will be run through                          */
/* MaxBL The maximum number of iterations for local
search    */
/* ec  The compactness epsilon                          */
/* eh  The homogeneity epsilon                          */
/* Maximi  The set of maximals (i-th maximal)
Begin
kNeighborhood←Generate a random number between 1 and
n ;
CurrentSol←Generate a random solution that is found in the
neighborhoodkNeighborhood ;
CurrentCompCost←getCompCost(CurrentSol) ;
CurrentHomCost←getHomCost(CurrentSol) ;
UpdateMaximals(CurrentCompCost,CurrentHomCo
cont←1 ;
/* The set of maximals is the empty set (initially)     */
Maxim←Ø;
whilecont<MaxV NS do
   kNeighborhood←1 ;
   whilekNeighborhood ≠n do
   SolCand←Generate a random solution that is found in the
   neighborhoodkNeighborhood ;   /* Local search */
   NumLoops←0 ;
   BetterSol←SolCand ;
   costCandComp←getCompCost(BetterSol) ;
   costCandHom←getHomCost(BetterSol) ;
   UpdateMaximals(costCandComp ,costCandHom ) ;
         whileNumLoops<MaxBL do
         SolCandBL←Generate random neighbor solution of
         BetterSol ;
         costCandBLComp←getCompCost(SolCandBL) ;
         costCandBLHom←getHomCost(SolCandBL) ;
   UpdateMaximals(costCandBLComp ,costCandBLHom ) ;
   if (costCandBLHom<costCandHom
   AndcostCandBLComp≤costCandComp )
   Or(costCandBLHom≤ costCandHom
   AndcostCandBLComp<costCandComp)
   Or(|costCandBLHom –costCandHom|<eh
   And|costCandBLComp –costCandComp|<ec)
   then   BetterSol←SolCandBL;
         costCandComp←costCandBLComp ;
         costCandHom←costCandBLHom ;
         NumLoops←MaxBL + 1 ;
  else
  NumLoops←NumLoops + 1 ;
  end
end
         SolCand←BetterSol ;       /* End local search*/
```

```
   if (costCandHom<CurrentHomCost And
costCandComp≤CurrentCompCost)
Or(costCandHom≤CurrentHomCost And
costCandComp<CurrentCompCost) Or(|costCandHom –
CurrentHomCostj|< ehAnd |costCandComp–
CurrentCompCost|<ec) then
CurrentSol←SolCand ;
   CurrentCompCost←costCandComp ;
   CurrentHomCost←costCandHom ;
   else
   kNeighborhood←kNeighborhood + 1 ;
end
end
cont←cont+1 ;
end
Return CurrentSol;
```

Algorithm 7. Compactness and homogeneity with Pareto and maximals condition

```
/* It checks if the pair (Comp; Hom) is a current maximal, in
that case it is added to the set of Maximals Maxim and the
pair dominated by this is removed from the set (if it exists)*/
/* if the set of Maximals is not empty*/
if |Maxim| > 0 then
/* for each element of the set of maximals    */
for i ←1 to |Maxim| do
(CompMax, HomMax) ←Maxim_i ;
/* If they are comparable and the pair (Comp,Hom)
dominates (CompMax,HomMax)*/
if (Hom<HomMax And Comp ≤CompMax) Or
(Hom≤HomMax AndComp <CompMax) then

/* (Comp,Hom) is the maximal until this moment and
replaces(CompMax,HomMax)*/
/*Remove(CompMax,CompHom) fromthe set of maximals*/

        Maxim_i←Maxim_i-(CompMax,CompHom)
/* Add (Comp,Hom) to the set of maximals*/
        Maxim_i←Maxim_i∪(Comp,Hom);   else
/*   If   they   are   comparable   and   the   pair
(CompMax,HomMax)dominates (Comp,Hom)        */

ifHomMax<Hom And CompMax≤ Comp) Or
(HomMax≤Hom And CompMax< Comp) then
/* (CompMax,HomMax) remains as the maximal and
(Comp,Hom)is discarded*/
Exit the function;
end
/* if they are not comparable, keep comparing with the rest
ofthe current maximals     */
end
end
end
/* If this step is reached it means that (Comp; Hom) is not
comparablewith any of the current maximals or that the set of
maximals iscurrently empty, therefore it becomes a maximal
        */
Maxim_i←Maxim_i  (Comp,Hom);
```

Algorithm 10. Function U pdateMaximals(Comp,Hom)

B. Tests

The pre-processing of spatial and census data, is common when it is important to retrieve relevant information for the problem to solve. For MC, the census data correspond to a data base of the INEGI census from the year 2000 [23] currently these databases are not available anymore due to the update of the 2010 census. These data are known as AGEBs. Example: Assume that there is a government program to attend education issues for the underage masculine population, in such a way that the distribution of this sector of the population must be known. Suppose that an expert of the population problem asks for groupings of 2, 8, 32, 64 and 100 groups having 469 Agebs to process. In accordance to things described in previous sections, the answer to a problem as the one we have defined starts with the selection of the variables of interest. These variables have the nomenclature X + natural number or Z + natural number and they are retrieved from a database to form the matrix of partitioning for homogeneity (see equation 12). The SQL query to choose the variables that take part in the grouping is:

```
censo15.mdb

SELECT id AS Ageb, Z002 AsVar from cdata
WHERE (x001 BETWEEN (SELECT MAX(x001)
FROM cdata) * 0 / 100 AND (SELECT MAX(x001) FROM
cdata) * 100 / 100) AND (x003 BETWEEN (SELECT
MAX(x003) FROM cdata) * 0 / 100 AND (SELECT
MAX(x003) FROM cdata) * 100 / 100) AND (x007
BETWEEN (SELECT MAX(x007) FROM cdata) * 0 / 100
AND (SELECT MAX(x007) FROM cdata) * 100 / 100)
```

TABLE I POPULATION VARIABLES

AGEB	Var
x001	Masculine population under 6 years old
x003	Masculine populations between 6 and 11 years old
x007	Masculine population from 15 to 17 years old
z002	Masculine population (Homogeneity Variable)

Our MC algorithm starts with a random solution in accordance to VNS. It is estimated that at least 10 runs for each instance must be done to trust an approximated result. However, each solution returns a set of minimal solutions that from the PF. The minimals from each run have been concentrated in a list to get a final set of minimals with Nodom[17]. For illustrative purposes Table 0 presents 5 runs for the instance of 2 groups with VNS parameters of 15 for local search and 2 for neighborhood structures. The CPU used is a dual core AMD of 2.0 Ghz and 2 GB in RAM.

TABLE II FINAL RESULTS FOR FIVE RUNS OVER THE INSTANCE OF TWO GROUPS

Test number for two groups	Time (seconds)	Minimals	
1	180	4651927	5830
		4027761	7586

		3998703	35908
		8418672	1134
		3860558	203380
2	191	3896887	108510
		4062652	3962
		5221435	1090
		3896898	20122
3	199	4274664	5038
		3911128	101652
		3962195	101368
		3861364	210136
		4462969	2364
		4225945	7132
		3964625	39146
		4075477	18332
		3800188	313336
		4028417	33306
		3853377	298410
4	178	4060232	1012
		5098746	6
		3977408	1516
		3775415	84998
5	200	4099570	50674
		3939251	56028
		4132966	43278
		4193850	446
		3830369	81072

Finally, in Figure 1 the minimals for the 10 runs associated to each instance (2, 8, 32, 64 and 100 groups) are shown. It is estimated that the greater the number of groups the homogeneity value presents better approximation but the compactness remains relatively stable, result that must be proven.

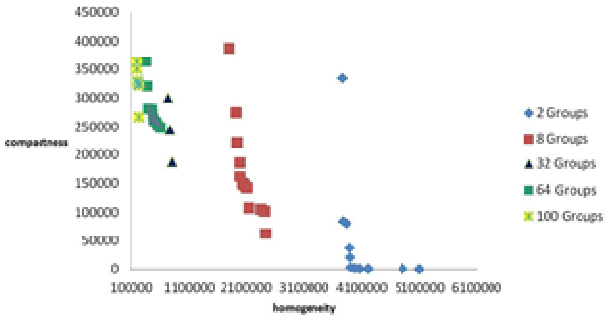

Fig. 1 Pareto frontier for 2, 8, 32, 64 and 100 groups

VI. CONCLUSIONS

Along this work we have presented a process of clustering development. The first proposal, reported with Simulated Annealing and VNS, is single-objective with optimization for geometric compactness. The second inherits the properties of partitioning to solve MC with relative weakness in the minimals due to the fact that the PF is obtained with 2 or even 4 additional dominated solutions, very close to the PF. The third proposal, our contribution, denominated MC is a partitioning around the medoids algorithm within a multiobjective context with VNS, it surpasses the previous algorithm despite it keeps central procedures. However, a diversification in the neighborhood structures has been implemented observing a Pareto Frontier with better distribution of the solutions along the frontier. Furthermore, our final algorithm has interesting properties because it achieves the Pareto frontier with a reliable method, analogous to the properties of the minimals in a Hasse Diagram, where these minimals are non-dominated and non-comparable points. The efficiency of the minimal solutions generated by our algorithm is confirmed when the generated solutions are ran through Nodom, an algorithm that filters non-dominated solutions. On the other hand, most of the grouping algorithms with only one optimized quality measure can work well but only for certain number of data or some lack robustness with respect to the variations in shape and uniformity of the cluster, furthermore the proximity to the optimum. In this work, we have proposed an alternative approach: optimizing simultaneously two objectives with VNS in a clustering problem for spatial data, however, our algorithm can group other kind of data. We have showed that with the approach described to solve MC, our method offers robustness in the selected solutions that form the Pareto Frontier but a problem that concerns many researchers in multiobjective is yet to be attended: the real Pareto Frontier.

REFERENCES

[1] Bação, F., Lobo, V. &Painho, M. (2005) Applying genetic algorithms to zone design, *SoftComput.*,**9** (6), 341–348.

[2] Kalcsics, J., Nickel, S. &Schröder, M. (2005) Toward a unified territorial design approach: Applications, algorithms, and GIS integration, *Top 13*,**1**, 1–56.

[3] Zoltners, A. &Sinha, P. (1983) Towards a unified territory alignment: A review and model, *Management Science*,**29**, 1237–1256.

[4] Brucker, P. (1977) On the complexity of clustering problems, *Optimization and Operations Research*,**157**, 45–54.

[5] Kaufman L., Rousseeuw P. (1997) Clustering by means of medoids, *Statistical Data Analysis based on the L1 Norm*, North-Holland, Amsterdam, 405-416.

[6] Hansen, P., Mladenovic, N. & Moreno, J.A. (2008) Variable neighbourhood search: methods and applications, *4OR*,**6** (4), 319–360.

[7] Mladenovic,N. & Hansen, P. (1997) Variable neighborhood search, *Computers & OR*,**24** (11), 1097–1100.

[8] Bailey, K.D. (1994) *Typologies and Taxonomies: An Introduction to Classification Techniques*, Sage Publications.

[9] Aldenderfer, M.S. & Blashfield, R.K. (1984) *Cluster analysis*. California Sage Publications, Inc.

[10] Piza, E. V. and Murillo, L. & Trejos, J. (1999) Nuevas tnicas de particionamiento en clasificación automática., *Revista de Matemica: Teor y Aplicaciones*,**6** (1).

[11] Bernábe, B., Espinosa, J.E., Ramírez, J. & Osorio., M.A. (2011) A Statistical comparative analysis of Simulated Annealing and Variable Neighborhood Search for the Geographic Clustering Problem, *Computación y Sistemas*,**14** (3), 295–308.

[12] Velmurugan, T. & Santhanam, T. (2010) Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points, *Journal of Computer Science*,**6** (3), 363–368.

[13] Pelta, D.A. (2000) *Algoritmos Heurísticos en Bioinformática*, Ph.D thesis, Universidad de Granada, Spain.

[14] Lara, A. (2003) *Un estudio de las Estrategias Evolutivas para problemas Multiobjetivo*, Masters thesis, Instituto Politécnico Nacional, México.

[15] Campbell, P.J. (1992) The origin of Zorn's Lemma, *Historia Mathematica*, **5** (1), 77–89.

[16] Gierz, G., Hofmann, K., Heinrich and Keimel, K., Lawson, J.D., Mislove, M. & Scott, D.S. (2003) *Continuous Lattices and Domains*. Cambridge University Press, England.

[17] Kung, H. T., Luccio, F. & Preparata F. P. (1975) On Finding the Maxima of a Set of Vectors, *J. ACM*,**22** (4), 469–476.

[18] Bernábe, B. & Guillén, C. (2012) Búsqueda de entorno variable multiobjetivo para resolver el problema de particionamiento de datos espaciales con características poblacionales, *Computación y Sistemas*,**16** (3), 335–347.

[19] Niemi, R.G., Grofman, B., Carlucci, C. & Hofeller, T. (1990) Measuring Compactness and the Role of a Compactness Standard in a Test for Partisan and Racial Gerrymandering, *The Journal of Politics*,**52** (4), 1155–1181.

[20] Young, H.P. (1988) Measuring the Compactness of Legislative Districts, *Legislative Studies Quarterly*,**13** (1) 105–115.

[21] Rincón, Eric. (2009) *Diseño de zonas geométricamente compactas utilizando celdas cuadradas*, Ph.D thesis, Universidad Nacional Autónoma de México.

[22] Insua, D. (1987) *Sobre soluciones optimas en problemas de optimizacion multiobjetivo, Trabajos de Investigación Operativa*,**2** (1), 49–67.

[23] *Censo General de Población y Vivienda, 2000.*Available from: http://www.inegi.org.mx/sistemas/microdatos2/default.aspx?c=14061&s=est