**Lecture Notes in Management Science** (2013) Vol. **5**: 57–66
5th International Conference on Applied Operational Research, Proceedings
© Tadbir Operational Research Group Ltd. All rights reserved. www.tadbir.ca

ISSN 2008-0050 (Print), ISSN 1927-0097 (Online)

# Clustering of characteristics over spatial data

María Beatriz Bernábe Loranca [1,2], Rogelio González Velázquez [1], Elías Olivares Benitez [2], David Pinto Avendaño [1], José Luis Martínez Flores [2] and J R Vanoye [2]

[1] *Benemérita Universidad Autónoma de Puebla,*
*Facultad de Ciencias de la Computación Puebla, México*
*dpinto@cs.buap*

[2] *Universidad Popular Autónoma del Estado de Puebla,*
*Posgrado de Logística y Dirección de la cadena de Suministro, México*
*elias.olivares@upaep.mx; joseluis.martinez01@upaep.mx*

**Abstract.** For problems that require a spatial data analysis, the use of statistical methods that take into account the geographical location or work with the descriptive characteristics of the data in the aggregation process is frequent. To solve a clustering problem over population data, considering the quantitative values of the variables that describe the data is necessary. In these cases, it is assumed that said variables have a high correlation, which suggests a statistical analysis with the goal to achieve a consistent subset of these variables. Even when we count with a subset of variables without redundancies and due to fact that specific problems of population character demand a reduced number of variables, a procedure of data selection under criteria boundaries is important; in this way achieving a subset of variables that describe a specific population problem is possible. Two objects are generated from this procedure: an associated distances matrix formed with the chosen variables and a vector of census-descriptive variables, which are processed by a partitioning algorithm with homogeneity restrictions for a population variable of interest. The homogeneity in the clustering of Agebs (Basic Geo-statistical Areas) is very useful due to the fact that balanced groups are wanted, with respect to a value of a census variable that responds to a population problem. The objective of this work resides in solving the spatial partitioning problem for geographic data under homogeneity restrictions where the variables to consider are directly related with the censuses in Mexico. The Agebs have a geographical composition of latitude-longitude and a vector of 167 descriptive variables of census kind. The partitioning problem is of combinatory character, such that the use of Variable Neighborhood Search (VNS) has been necessary to optimize the objective function with the associated homogeneity restriction. Finally the results are presented for a homogeneous grouping case for economically inactive population.

**Keywords:** compactness; homogeneity; partitioning; variable neighborhood search

## Introduction

Different geographical, spatial and population problems involve objects with a clear geographic component (their location in the space in relation to other objects is necessary for their study). Some examples of this kind of problems are: the distribution of a disease, the possible relationship between the income levels, electoral preferences and general problems of population character.

The problem of analyzing phenomena with a spatial component is similar to analyzing data that are related among them and determined statistical methods are employed for this purpose. The statistical concept of correlation between objects has an equivalent in geostatistics: the spatial autocorrelation; in it, the relative position of some observations with others can be important and must be explicitly included in the analysis.

To treat a problem where identifying associations between variables is important, a classic clustering like the statistical clustering isn't viable if we consider that is necessary to define quantifiable conditions of values in the variables to form groups. In this point, for the development of our work, we have decided to start the study with a correlations analysis and later applying a procedure to select variables. This means an even more exhaustive debugging that gives place to a clear description of the population problem. The next step is to group the partitions of the chosen data with a combinatory optimization method with the minimization of distances as cost function and an additional homogeneity restriction for a given population variable and the heuristic approximation method is VNS. In this way, the spatial data that will form the groups will be those that satisfy determined values of interest, being compact and homogeneous for a census-population variable. Properties from the classic partitioning have been considered for the implementation where the objective function of Euclidean distances minimization is optimized and implicitly the geometric compactness is solved. For the compactness the spatial location of the geographic data has been considered and for the handling of homogeneity over a specific variable, the census data.

In accordance to the above, the present work is organized as follows: section 1 as introduction. In section 2 the procedure to select the variables is exposed. The section 3 is occupied to present the model of partitioning with a homogeneity restriction. In section 4 a brief experiment is shown and finally in section 5 the conclusions are presented.

## Preprocessing of geographic-spatial data (Agebs)

The main purpose of the spatial data analysis is to detect and model the possible patterns that form the data. In distinct geographic problems, the data are found aggregated in areas as happens with the basic census unit for socioeconomic data, this means that the zones are geometrically irregular (the Agebs data that we treat in this article don't have a polygonal structure). In these terms, the analysis of aggregated data presents diverse limitations and difficulties due to the fact that generally the areas are administrative or legal units (municipalities, states, electoral districts, etc.). The population censuses constitute an example of data aggregated by area since the information is gathered home by home but is made public grouped in units of area. In Mexico, data are aggregated by Ageb, which are a partition of the municipalities and the criteria used for the creation of the Agebs are independent from those used for the electoral sections. As well is necessary to underline that the regionalization criteria (grouping of geographical zones or units), tend to be based on a single variable, for example number of habitants, which provokes that they aren't really adequate to analyze other factors such as income or educational level. Thus, even when a regionalization is made following a homogeneity criterion, is practically impossible that the same is valid for all the variables.

In this scenery is focused our work: the use of a statistical procedure and a criteria for the selection of variables to start the population study, which is called pre-processing of the data that produces a matrix adjusted to the selected variables and a vector of variables. These will be the input data to a partitioning algorithm maintaining geometric compactness with homogeneity restriction for a census variable.

## Correlations and Selection of Variables for Agebs

The geographical data under study present a high correlation that in some cases must be distinguished with the goal of reducing statistical redundancy. Other multi-variate statistical aspects have been applied to the Agebs and some groups of variables have been clearly identified, Bernábe (2004). In this section we present the correlated groups that contribute to the selection of variables to facilitate the description of a population problem with greater accuracy.

The association of variables in accordance to the nature of the problem is based on eliminating the high correlation between these. Three main principles were considered for the variables selection: 1) availability of the data, 2) data representation and 3) the much correlated group. In the table 1 a summary of these groups can be seen (correlation .97 to .99).

**Table 1.** Agebs' variables with high correlation.

| ***0.97*** | ***0.96*** | ***0.95*** | ***0.94*** | ***0.93*** | ***0.92*** | ***0.91*** | ***0.9*** |
|---|---|---|---|---|---|---|---|
| *x6,x7,x8* | *z50,z52* | *x2,x6* | *x6,z85,**z105**,z75* | *z46,z95* | *z126,z133* | *z85,z170* | *z51,z97,z116,z106* |
| *z72,z79,z116* | *z73,z107,z112* | *z46,z47* | *z95,z149* | *z118,z161* | | | *z71,z96* |
| *z73,z80,z146,z153,z159* | *z85,z148* | *z50,z73,**z104**,z117,z127* | | | | | *z110,z111* |
| *z82,z118* | *z126,z130* | *z51,z74* | | | | | *z85,z113* |
| *z108,z117* | | *z128,z126* | | | | | *z121,z123* |
| *z167,z170* | | | | | | | |

Z104 12 years old and over economically inactive population that study.
Z105 12 years old and over economically inactive population that do housework.
Z103 Unemployed population (correlation of .80)
Z102 Economically inactive population (low income employment profile with correlation of .85).

The table 2 shows a correlation of .85 and .80 where the variables Z102 and Z103 are located. The numbers 4, 3, 2, 1 and 0 tag the subsets of variables with a correlation of 0.85: 0 means that they are not related, 1 they are variables with dependence relationship and relationship children-women, 2 indicates low income employment profile, to 3 belong the variables with low income population profile and 4 are the people with high income coming from another entity. For the group with correlation 0.8, 0 are the unrelated variables and 1 the variables related to the population and 2 the variables unrelated to the population (indexes and averages), CVA means correlation of variables. The variables' description can be seen in (http://www.cs.buap.mx/~bety/InfCensal2.htm).

**Table 2.** Agebs variables with correlation .80 and .85

| *C VA* | *-0.85* | *-0.8* | *C VA* | *-0.85* | *-0.8* |
|---|---|---|---|---|---|
| *x013* | 0 | 0 | Z077 | 2 | 1 |
| *Z060* | 0 | 0 | Z085 | 2 | 1 |
| *Z083* | 0 | 0 | Z110 | 2 | 1 |
| *Z109* | 0 | 0 | Z167 | 2 | 1 |
| *Z114* | 0 | 0 | Z051 | 3 | 1 |
| *Z115* | 0 | 0 | Z070 | 3 | 1 |
| *Z137* | 0 | 0 | Z071 | 3 | 1 |
| *Z138* | 0 | 0 | Z121 | 3 | 1 |
| *Z142* | 0 | 0 | Z126 | 3 | 1 |
| *Z147* | 0 | 0 | Z055 | 4 | 1 |
| *Z046* | 0 | 1 | Z118 | 4 | 1 |
| *Z053* | 0 | 1 | Z100 | 0 | 2 |
| *Z093* | 0 | 1 | Z162 | 0 | 2 |
| *Z103* | 0 | 1 | Z163 | 0 | 2 |
| *Z106* | 0 | 1 | Z164 | 0 | 2 |
| *Z141* | 0 | 1 | Z048 | 1 | 2 |
| *Z150* | 0 | 1 | Z049 | 1 | 2 |
| *Z050* | 2 | 1 | Z102 | 2 | 1 |

## Procedure for the Selection of Variables

We have mentioned that when the spatial and census data are processed, a statistics exploratory analysis for their adequacy is required; this allows the possibility to grant quality to the final data and to make them accessible to different subsequent treatments. In previous works, multivariate techniques have been employed identifying important relationships; however, for the goals of this work, a summarized correlations process has been presented.

The treated data have been defined in a descriptive way by a vector of 171 variables. The empirical evidence that is provided for the clustering is based on the 469 Agebs from the metropolitan zone of the Toluca valley ZMVT. The partitioning over spatial data with a homogeneity restriction has inherited the properties of compactness Bernabe (2011) and can be seen as application software with a modular structure of procedures Bernábe (2011a). For this work, the compact partitioning only requires the dissimilarity matrix as input and the descriptive information of the census variables is unnecessary Bernabe (2011). The choosing of variables is important when homogeneity is calculated to a partition over compactness as it is the objective of this work. For the homogeneous clustering over variables that we expose here, two steps are required for integrating the distances matrix and the set of selected variables in order for them to work as input in the algorithm: 1) A selection process of adequate variables for a particular population problem. The result is a subset of variables that can be restricted in their census values. The subset of variables is obtained through a set of SQL queries Bernábe (2010). From the resulting group of variables of the previous step, an adjusted dissimilarity matrix is obtained for the selected variables. This means that we count with a distances matrix over specific variables. The traditional classification algorithms don't condition the variables and for diverse problems is useful to restrict the variables as well as the values of these. For example, if it is wished to obtain a partition of economically inactive women, only those variables of women related to economical inactivity are needed and with a certain percentage of these to homogenize one of the chosen variables. In this sense, a categorization process has been implemented that starts with the extraction of population variables with a search procedure in the data database of variables, thus providing a set of Agebs that meet the specified characteristics Bernábe (2010).

## The model proposed

The compact-homogeneous partitioning algorithm for variables considers in the clustering a subset of zones enclosed in their descriptive values. A compact partition is obtained and the group's homogeneity is calculated for a chosen variable. The combinatory model of this partitioning problem for zones of Agebs is binary mixed-integer with geometric compactness and homogeneity restrictions. The clustering of Agebs is solved in such a way that the Agebs that form the groups are geographically very close to each other, for this, the cost function minimizes the distances between them. Informally, the strategy is based on randomly choosing Agebs as centroids that determine the number of groups. Those Agebs that aren't centroids that have the shortest distance to a determined Ageb-centroid are the members of a group. Given the combinatory character of the partitioning, is an NP hard problem, Trejos et al (1990), and is necessary to include heuristics in the partitioning algorithm to minimize the objective function with a homogeneity restriction on the census variable. For the handling of the computing cost the Variable Neighborhood Search has been chosen given its high capability to escape local optima Mladenović N and P. Hansen (1997).

The following model has been implemented using the basic properties of partitioning algorithms, where implicitly the necessary restrictions for the fulfillment of geometric compactness are satisfied. In this work we have added an additional homogeneity restriction for population variables, which is of great use for clustering where territory design problems of population character must be solved.

### Homogeneity-Compactness model

The combinatorial optimization model of Territory Design Problem is:
Let $\mathrm{T}=\{\mathrm{BGU}_1,\mathrm{BGU}_2,\ldots\mathrm{BGU}_\mathrm{n}\}\subseteq \mathrm{R}^2$ a territory of basic geographical statistical areas BGAs. Let $C=\{c_1,c_2,\ldots,c_n\}$ be the set of centroids where $\mathrm{c}_\mathrm{i}\in\mathrm{BGU}_\mathrm{i}$ (basic geographical units) and $c_i=(x_i,y_i)\ \forall i=1,2,..,n$ with x-longitude and y-latitude. Let $d_{ij}=d(c_i,c_j)$ Euclidean distance from centroid *i* to *j*. let $G_i$ a subset of $C$ called group. BGAs are indivisible units and belong to groups. Groups are represented by centroids. The problem of partitioning T is to find a partitioning $P=\{G_1,G_2,\ldots,G_p\}$ so that $\min z=\sum_{i=1}^{p}\sum_{j\in G_i} d_{ij}$ *(1, compactness*).
The constraints are:

$$G_i\neq\varnothing,\ \forall i=1,2,..,p;\ G_i\cap G_j=\varnothing,\ \forall i\neq j;\qquad \bigcup_{i=1}^{p}G_i=T.$$

The minimization of z ensures the compactness, but in this case, it's necessary that the groups be as homogeneous as possible considering a population variable of interest. The model in question is binary mixed-integer and makes use of the binary variables for models of this kind. Additionally with the goal of modeling the groups' homogeneity with respect to the variables of attributes of the population we must include in the mathematical modeling the following considerations:

$A$ *the set of quantifiable attributes from which a subset of variables will be selected in accordance to the specific problem to be treated.*

$\mathrm{X}_{ij}=1$ *if the i^th^ centroid is in the j^th^ group and 0 otherwise*

$\mathrm{VA}_{kj}$ *is the value of the k^th^ attribute belonging to the j^th^ BGU*

$\alpha_k$ , $\beta_k$ *values of the tolerance parameters for*

$$\alpha_k\leq\mathrm{VA}_{kj}\leq\beta_k\qquad\forall k=1,\ldots,|\mathrm{A}|\quad y\quad\forall j=1,\ldots,n.$$

$\mathrm{VA}_{ki}=\sum_{j=1}^{|G_i|}\mathrm{VA}_{kj}\mathrm{X}_{ij}\quad\forall k=1,..,.|\mathrm{A}|.$ *is the value for the k^th^ G*

$\overline{\mathrm{VA}}_k=\dfrac{\overset{n}{\underset{j=1}{\text{å}}}\ \mathrm{VA}_{kj}}{|\mathrm{P}|}$ *is the goal value for the k^th^ attribute in any BGU.*

In the implementation of the methodology is intended for the partition P to be homogeneous with respect to one attribute therefore we must minimize the difference between the goal value $\overline{\mathrm{VA}k}$ and the value for each group $\mathrm{VA}_{ki}$. If we define the homogeneity norm of the partition P as the $\|P_H\| = \max\left\{\overline{\mathrm{VA}_k} \text{-} \mathrm{VA}_{ki}\right\}" k, i.$, then the ideal homogeneity is when

$$\lim_{\|P_H\| \text{®} 0} \overset{|G_i|}{\underset{j=1}{\text{å}}} d_{ij} \mathrm{X}_{ij} \quad (\textit{2, homogeneity}).$$

The description of the problem suggested in the previous model indicates that heuristic methods must be employed to obtain a satisfying solution with a reasonable computing cost. VNS is an efficient search metaheuristic and we have statistically proven that VNS responds better to the optimization problem by partitions than SA, Bernábe (2011) and Mladenović N. and P. Hansen (1997). Basic VNS is a strategy that alternates local search with random movements in the neighborhood structures which vary in a systematic manner. The steps for basic VNS can be seen in Bernábe (2011a). Two parameters are of interest in VNS: the neighborhood structures (NS) and local search (LS).

**Clustering with a VNS algorithm and a homogeneity restriction**

In this work, the compact partitioning with VNS that achieves an approximated compact solution is similar to the one exposed in Bernabe (2011). However, a subroutine has been incorporated with the goal of achieving a homogeneity cost of the compact solutions. This algorithm is extensive, but it can be reduced to 2 general facts:

```
1 Obtains a random initial compact solution with VNS minimizing the implicit objective in equation (1, compactness). The cost is stored in getCosteComp(Sol)←cost
2 Calculates the homogeneity cost with the following procedure employing equation (2, homogeneity)
Function getCosteHom (Sol)
total←0
cost←0
For i←1 to n do
      ng←Get the number of group to which AGEB i belongs
      total←total + $Val_i$
      $totalGroup_{ng}$←$Val_i$
end For
idealAverage←total/k
For j←1 to k do
      cost←cost + |totalGroupj - idealAverage|
End For
getCostHom(Sol)←cost
```

## Experimental results

The problem that has been considered takes 4 related variables with economically inactive population. Assuming that a government program exists to push support actions to this population sector, a set of congruent groupings is required that represent the distribution of this population with respect to the following census variables, 1) 12 years old and over economically inactive population that study, 2) Unemployed population, 3) 12 years old and over economically inactive population that do housework and 4) Economically inactive population.

In a purely experimental way a set of 15 test runs have been chosen. Homogeneity is kept in economically inactive population Z102. The results can be seen in table 3, which provides information to develop a factorial statistical experiment later and estimate or calibrate the best parameters for this problem.

In the table the notation is: NS Neighborhood Search, LS Local Search, *t* computing cost and *HC* Homogeneity Cost.

**Tabla 3.** Random experiment for economically inactive population.

| *test* | *solutions* | *iterations* | *NS* | *LS* | *t* | *HC* |
|---|---|---|---|---|---|---|
| 1 | 4 | 2897 | 4 | 30 | 14 | 447830 |
| 2 | 6 | 911 | 2 | 15 | 4 | 448993 |
| 3 | 6 | 1877 | 3 | 20 | 49 | 451559 |
| 4 | 8 | 6713 | 8 | 40 | 32 | 449746 |
| 5 | 12 | 18123 | 20 | 75 | 96 | 446683 |
| 6 | 6 | 188745 | 200 | 1500 | 3318 | 445137 |
| 7 | 19 | 1901120 | 2000 | 15000 | 4294 | 443896 |
| 8 | 6 | 904 | 2 | 15 | 4 | 451807 |
| 9 | 11 | 17862 | 20 | 150 | 157 | 447447 |
| 10 | 8 | 4619 | 6 | 35 | 22 | 448057 |
| 11 | 12 | 1899389 | 2000 | 15000 | 3820 | 443826 |
| 12 | 6 | 1062 | 2 | 15 | 3 | 451965 |
| 13 | 7 | 990 | 2 | 15 | 4 | 448323 |
| 14 | 15 | 27395 | 30 | 250 | 15 | 448427 |
| 15 | 4 | 27447 | 30 | 250 | 15 | 448150 |

The best homogeneity cost happens when the search reaches the smaller value and corresponds to the test run 11 with 12 solutions found with the highest parameters, this result can be seen in the following figure 1.

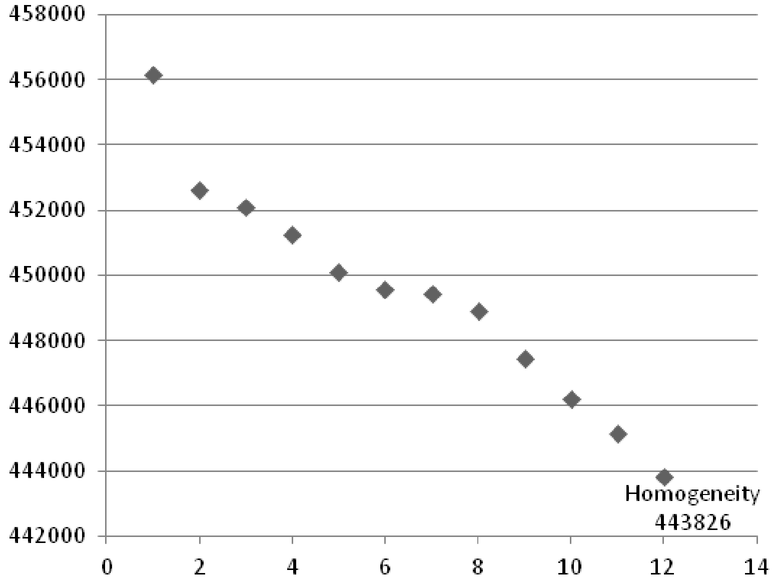

**Fig. 1.** Behavior of the Homogeneity cost of VNS for test run 11.

## Conclusions

The strategy used to choose variables and obtain a distances matrix associated to the subset of variables, allows to count with reduced data to respond in an adequate way to the description of a population problem, in this way the homogeneity-compactness clustering will generate geometrically compact groups but homogeneous for a descriptive variable of the problem it deals with. However, is necessary to review the generated solutions with a geographic information system and examine the map of the zone with block cartography to study the distribution percentage of the population chosen.

On the other hand, the method that has been presented responds very well to the partitioning under the calculus of homogeneity and provides a solution with a very good approximation with the heuristic that has been used.

The integration of the creation of the map for the best solution found hasn't been reported in this work, but the inclusion in the system has already given partial results. In this point, the problem that we have presented is of Territory Design kind and a solution approach in this direction must be presented where the implications of the results are included in a map referenced with the territorial partition, Kalcsics et al (2006) and Tavares et al (2007).

Finally, the solution that best responds to the value of homogeneity is the test run 11 with 12 accepted solutions, 1899389 iterations, 2000 neighborhood structures and 15000 local search iterations. The computing time was 3820 seconds and the homogeneity cost was calculated in 443826 seconds.

## References

Bernábe LB. & Sales, LR (2004). Application of Non-Supervised Classification to Population Data. ICEEE/CIE 2004, International Conference on Electrical and Electronics Engineering, Acapulco México, ISBN 0-7803-8531-4.

Bernábe LB., Espinosa RJ., Ramírez RJ., Osorio LM. (2011). A Statistical comparative analysis of Simulated Annealing and Variable Neighborhood Search for the Geographical Clustering Problem. Computación y Sistemas, 14 3: 295-308.

Beatriz Bernábe Loranca (2010). Desarrollo de un modelo para la determinación de Zonificación Óptima. Tesis Doctoral, Investigación de Operaciones, UNAM.

Kalcsics, J., Nickel, S., Schröder, M. (2005). Toward a unified territorial design approach: Applications, algorithms, and GIS integration, TOP 13 (1) 1–56.

Mladenović N. and P. Hansen (1997). Variable neighborhood search. Computers & Operations Research, 24:1097-1100.

Tavares P., Figueira F., Mousseau J., Roy V. (2007). Multiple criteria districting problems.The public transportation network pricing system of the Paris region. Annals of Operations Research, 154, pp. 69-97.

Trejos Z. J., Castillo E.J., González V. (1990). Análisis multivariado de datos: Métodos y Aplicaciones Escuela de Matemática Universidad de Costa Rica (In Spanish).