Fuzzy geo-processing for characterization of social groups: an application to a Brazilian mid-size city

G. R. A. Gonzalez[1], A. G. Evsukoff[1], R. C. Pinto[1], A. P. B. Sobral[2] & J. A. Silva[2]
[1]*Civil Engineering Department,*
Federal University of Rio de Janeiro – COPPE/UFRJ, Brazil
[2]*Department of Statistics,*
Federal University of Juiz de Fora – UFJF, Brazil

Abstract

This paper presents a method for the spatial representation of social-economic groups. This work is based on the Brazilian census geo-referenced data for the revenue and education level of 540 districts in a mid-size city. The data was analyzed by k-means clustering algorithms for determination of groups of similar behavior based only on the revenue and education level data. The groups were then plotted into the city map using geo-referenced information. The aim of this study is to analyze the spatial distribution of groups of equivalent socio-economic levels, taking into account the uncertainty of the classification process. The results show that the model is able to represent the distribution of the social groups in an inter-related and continuous space.
Keywords: cluster analysis, fuzzy classification, spatial data mining, socio-economic studies.

1 Introduction

Spatial data mining is a very attractive research area, since it provides a way to deal with spatial relationships found in data. Spatial data mining is useful in several industries when geo-referenced information must be taken into account such as government, marketing, oil and gas exploration [4]. In government, spatial data mining can be used to guide social politics and investments.


WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060051

The application described in this work has used census database with information of 540 sectors of the city of *Juiz de Fora*, a middle size Brazilian city. The city has a population of about 450,000 inhabitants and is located in a mountainous region in *Minas Gerais* state. The economy is based on industry, such as the automobilist industry and service companies.

Each record in the database represents a district sector, which is localized by the Universal Transverse Mercator (UTM) geographical coordinates and presents social and economic information about the district.

For this study, only the revenue and educational level information has been used, represented as ranges of average values. For each sector, the number of individuals on each range was recorded, as well as the total number of inhabitants living in the district.

This presents a method for spatial representation of group extracted from the database by a clustering algorithm. The methodology is sketched in Figure 1 the k-means clustering algorithm is used to identify groups based on education and revenue information. Fuzzy classification is then used to localize the groups in the city map using geo-referenced information. A grid is then generated to plot the fuzzy rules' model into UTM coordinates.

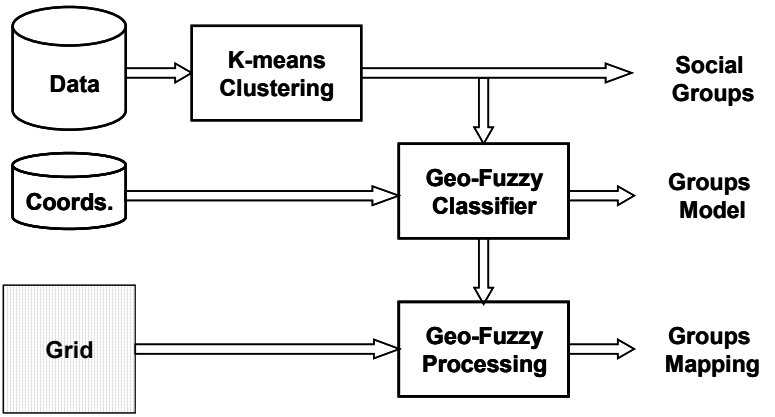

Figure 1: Fuzzy geo-processing method.

The remainder of the paper is organized as follows: next section reviews the basic issues of k-means algorithm in order to introduce the notation. Section 3 presents the fuzzy reasoning method for spatial modeling. Section four presents results and discussion. Finally some conclusions are drawn.

2 The k-means clustering analysis

Clustering is one of the most usual tasks in the process of data mining, helping the discovery and identification of distributions and patterns of interest. Data clustering is being used in several data intensive applications, including image classification, document retrieval and customer segmentation (among others).

Consider data set $\{(\mathbf{x}(t)), t = 1..N\}$, where each record contains the variables vector $\mathbf{x}(t)$. The well known k-means clustering algorithm aims to find a partition of the domain into a set of K clusters $\{C_1 \ldots C_K\}$, where each cluster C_i is represented by its center's coordinates vector $\mathbf{w}_i$.

The k-clustering algorithm assigns each record to the nearest cluster according to a distance measure (generally the Euclidean distance, as in this work). The algorithm iterates until reassigning points produces no changes [1]. The number of clusters is assumed to be fixed in k-means clustering, and must be given at initialization. The iterative algorithm minimizes the compactness of clusters, represented by the error function:

$$J = \sum_{t=1}^{N} \sum_{j=1}^{K} \left(\mathbf{x}(t) - \mathbf{w}_j\right)^2 \tag{1}$$

The k-means algorithms results on a partition matrix $\mathbf{V}$, containing one binary vector $\mathbf{v}(t) = (v_1(t) \ldots v_K(t))$ for each record, where:

$$v_j(t) = \begin{cases} = 0, \ if \ \mathbf{x}(t) \in C_j \\ = 1, \ if \ \mathbf{x}(t) \notin C_j \end{cases} \tag{2}$$

As the number of clusters must be given at initialization, the k-means clustering algorithm is generally executed for a range of number of clusters and a validation metric is used to validate clustering. Moreover initialization of clusters centers is provided randomly such that the solution may fall in local minimum. It is generally desired that, even for the same number of cluster, the algorithm be to execute a number of times to verify the stability of the solution.

3 Fuzzy spatial classifier

Fuzzy set approaches have been applied in the database and information retrieval areas for nearly 30 years. Applications to areas such as data mining and geographical information systems have been developed under the fuzzy logic theory [5].

Each coordinate $y_i(t)$ is described by a fuzzy partition $\mathbf{A}_i = \{A_{i1}, \ldots, A_{in}\}$ where $A_{ij} \in \mathbf{A}_i$ is a fuzzy set [2]. The number of fuzzy set for each coordinate is the same to simplify the computations. The coordinates were calculated by the geometric centroids of each district.

Strong normalized and triangular fuzzy partitions are used to represent each input variable. Trapezoidal membership functions are used for the two fuzzy sets at each end of the domain, as shown in Figure 2, to deal with off-limit points.
The fuzzy rules base relates the input fuzzy sets to the classes, in rules as:

$$if \ \mathbf{y}(t) \ is \ B_k \ then \ class \ is \ C_j \ with \ cf = \varphi_{kj} \tag{3}$$

The fuzzy set B_k in rule represents the combination of the fuzzy sets in the partition of each coordinates and defines a geographic region. For a given an

input $\mathbf{y}(t) = (y_1(t), y_2(t))$, all the combinations of fuzzy sets in each fuzzy partition must be considered in such a way that the model is complete, *i.e.* it produces an output for any input values. In Figure 2, the combination of two fuzzy partitions of 5 fuzzy sets each is shown.

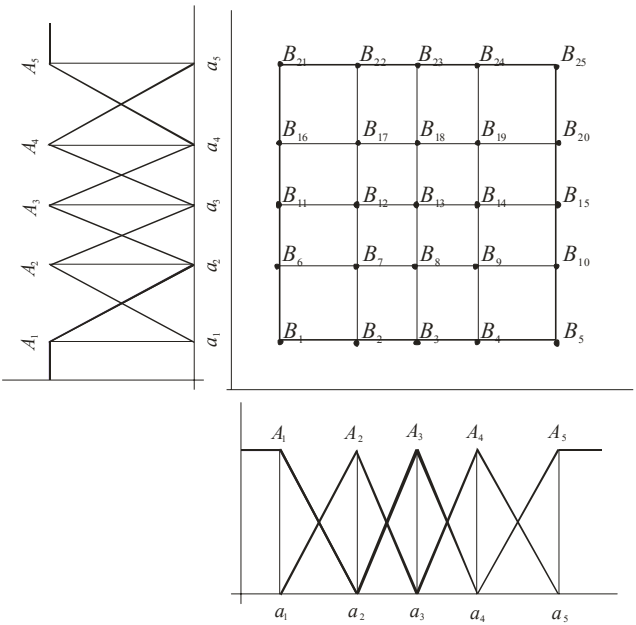

Figure 2: Construction of the geographic fuzzy set.

Each component $u_k(t) = \mu_{B_k}(t)$ of the fuzzification vector $\mathbf{u}(t)$ is computed as:

$$u_k(t) = \mu_{A_{1i}}(y_1(t))\mu_{A_{2j}}(y_2(t)),\ i, j = 1...n. \tag{4}$$

The confidence factor $\varphi_{kj} \in [0,1]$ in rule (3) represents the certainty of the rule. The confidence factor weight all rules in the fuzzy rule base. The value φ_{kj} represents how much the term B_k is related to the class C_j in the model described by the rule base.

The rule base can be represented by the matrix $\Phi = [\varphi_{kj}]$, of which each line is related to a geographic fuzzy sets B_k and each column is related to a class C_j.

The rule base weights are the kernel of the model described by the fuzzy rules (3) and its determination is calculated as described next.

3.1 Rule base identification

The rule base weight are computed from a data set T', where each sample $t = 1..N$ is a pair $(\mathbf{y}(t), \mathbf{v}(t))$, of which $\mathbf{y}(t)$ is the coordinates vector and $\mathbf{v}(t) = (v_1(t) \ldots v_K(t))$ is a row in the partition matrix computed by the k-means algorithm. Each sample t in the data set T' is related to socio-economic characteristics of the population data in the sample t of the data set T.

Each rule in the rule base is a sub-model that assigns a class (computed in k means cluster analysis) to the corresponding region of the domain. Each rule base weight φ_{kj} can be seen as a measure of how frequent the class C_j occurs in the region B_k. Under this interpretation, the rule base weights are computed as:

$$\varphi_{kj} = \frac{\sum_{t=1..N} u_k(t) v_j(t)}{\sum_{t=1..N} u_k(t)} \tag{5}$$

where $u_k(t)$ is the membership of the register t to the geographic fuzzy set B_k, computed as and $v_j(t) = \mu_{C_j}(\mathbf{x}(t))$, i.e. the membership of the social economic characteristics data in the register t to the cluster C_j.

3.2 Fuzzy spatial interpolation

In the third and last step of the methodology, a grid of testing points $\mathbf{z}(t) = (z_1(t), z_2(t))$ is generated to create a map of the clusters of concentration data into the geographic domain. The grid is represented by a testing set $T'' = \{(\mathbf{z}(t)), t = 1..M\}$, where M is the number of registers in it.

The fuzzy spatial interpolation aims to calculate the output of the fuzzy spatial classifier to each point of the grid, i.e. the class membership vector $\hat{\mathbf{v}}(t) = (\mu_{C_1}(\mathbf{z}(t)), \ldots, \mu_{C_K}(\mathbf{z}(t)))$, where $\mu_{C_j}(\mathbf{z}(t))$ is the output membership value of the grid coordinates $\mathbf{z}(t)$ to the class C_j.

The fuzzification vector $\hat{\mathbf{u}}(t)$ is computed for every point in the grid. Each component of the fuzzification vector is computed as product of the membership of each coordinate value to the respective fuzzy partition:

$$\hat{u}_k(t) = \mu_{A_{1i}}(z_1(t)) \mu_{A_{2j}}(z_2(t)),\ i, j = 1 \ldots n. \tag{6}$$

The class membership vector is computed from the input membership vector $\hat{\mathbf{u}}(t)$ and the rule base weights matrix Φ. Using the sum-product composition operator for the fuzzy inference, the class membership vector $\hat{\mathbf{v}}(t)$ can be computed as a standard vector matrix product as:

$$\hat{\mathbf{v}}(t) = \hat{\mathbf{u}}(t).\Phi \tag{7}$$

The number of fuzzy sets as well as the number of points in the grid controls the accuracy of the map.

4 Results and discussion

Only educational and revenue information, for each district sector, was used in this application. Each variable presents the number of inhabitants living in the district, was divided in six ranges. The education level ranges vary from non-alphabetized to graduate. The revenue ranges vary according to the lowest income and the highest income registered.

The k-means algorithm was applied to the data set and $K = 3$ clusters were found. The centers of clusters according to educational and revenue levels are shown in Figure 3. The Cluster 1 groups sectors with high revenue and education level population. Cluster 2 presents high concentration of middle revenue value and low educational grades. In Cluster 3 groups population with middle to high revenue values and low to middle educational grades.

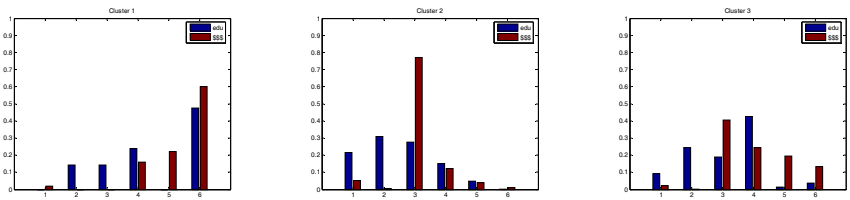

Figure 3: Clusters centers.

The support of each cluster, computed as the percentage of districts in each cluster, is shown in Table 1. Cluster 1 is represented by only one sector, it is considered an elite district that has the best revenue and education level. This result reproduced a typical Brazilian social class distribution, where a small portion of the population has access to higher education levels and revenue. The majority of the population is represented by Cluster 2, 69.81% of sectors are labeled on this one. The cluster 3 appears with 30% of the sectors, this clusters appears as the mid-class population.

Table 1: Clusters supports.

	Cluster 1	Cluster 2	Cluster 3
%	0.18	69.81	30.00
number	1	377	162

The fuzzy spatial classification, as presented above, allows the clusters to be mapped disregarding the sectors' boundaries. Uncertainty is taken into account as color codes computed by defuzzification of the classification results. The maps computed for each one of the clusters is shown in Figure 4 where the color

grades are proportional to the membership values of the districts to each cluster, which represents the uncertainty in the classification.

The Cluster 1, presented by the first map, is located on north of the city. This cluster appears in a small and concentrated area as it represents only one sector in the database.

The Cluster 2 spreads for most of the city area, and represents the majority of the population with low education level and average revenue. The area shown on the map is predominantly on the periphery of the city, which is another standard population distribution in urban areas characterized by the poorest inhabitants located on the surround area of the city.

The Cluster 3, represents sectors with average life quality level, is located in a discontinued area on south area typically known as commercial districts including downtown. The population on these regions is very heterogeneous especially at the oldest districts on south.

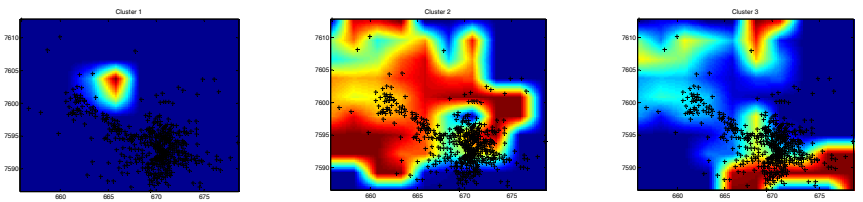

Figure 4: Clusters representation.

The result discussed above coincides with the actual population distribution of the *Juiz de Fora* city. The maps shown in Figure 4 can help to evaluate government actions and improve the selection of areas for investments.

5 Conclusion

The methodology proposed by this work has shown a good alternative to the traditional spatial classification models. The representation based in delimited regions of the geographical space does not allow visualizing the different classes beyond the sectors borders. This model presented a result closer to the reality, being capable to represent the distribution of the social groups in a continuous and inter-related space.

The limitation of this approach is given by using the geometric centers to characterize the whole district, especially on sectors of large areas; those sectors might be badly represented when the geometric centers do not match the population concentration center. Despite this limitation, most of sectors are small and dense, and the geometric centers can be assumed.

Acknowledgements

This work is supported by the Brazilian research agencies, CNPq, CAPES and FINEP.

References

[1] Duda, R. O., Hart, P. E., & Stork, D. G. (2000). "Pattern classification", 2nd ed. Wiley.

[2] L. A. Zadeh. *Fuzzy sets as a basis for a theory of possibility*. Fuzzy Sets and Systems, 1(1):3--28, 1965.

[3] Bezdek, C. J. e Pal, S. K. "Fuzzy Models for Pattern Recognition". IEEE Press, New York, 1992.

[4] A. G. Evsukoff, F. T. T. Gonçalves, R. P. Bedregal, N. F. F. Ebecken, "Fuzzy Classification of Surface Geochemistry Data Applied to the Determination of HC Anomalies".

[5] Patrick Bosc, Donald Kraft and Fred Petry. "Fuzzy sets in database and information systems: Status and opportunities" Fuzzy Sets and Systems, Volume 156, Issue 3, 16 December 2005, Pages 418-426.