Fuzzy Classification of Geodemographic Data Using Self-Organizing Maps

Miguel Loureiro[1], Fernando Bação[1], Victor Lobo[1,2]
[1] ISEGI – Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
{mloureiro, bacao, vlobo}@isegi.unl.pt
[2] Escola Naval - Alfeite, 2800 Almada, Portugal

Introduction

The convergence of geographical information systems and census data made available an immense volume of digital geo-referenced data. This has created opportunities for developing an improved understanding of a number of socio-economic phenomena that are at the heart of human geography. In spite of the potential of geodemographics, a number of problems which plague it have been described in the relevant literature (Birkin et al., 1998; Feng et al., 1998; Harris, 1998; Openshaw et al., 1994). These include the definition of the number of clusters; the variability in the size of Enumeration Districts (EDs) which influences the precision and resolution of data, and emphasizes the need to use classification algorithms which are robust to outliers; and the inherent fuziness of the geodemographic analysis.

We propose the use of Self-Organizing Maps (SOM) (*e.g.* Kohonen, 2001) as an adequate tool for geodemographics, not only because they are powerful and reliable algorithms for clustering (Bacao et al., 2005), but also because they provide effective visualization/exploration tools. Additionally we implement fuzzy membership functions, based on the SOM, which allow for productive "what if" analysis.

Self-Organizing Maps and Fuzzy Classification method

A SOM is an unsupervised neural network and has been used as visualization tool for high dimensional data. When a SOM is trained with a given dataset, its units tend to spread themselves in input space in a way that is proportional to some function of the density of the data patterns (Cottrell et al., 1998; Fort, 2005).

The SOM may also be regarded as a graph in the input space. We introduce a new way of calculating fuzzy memberships based on the

computation of variations in SOM unit density, which are proportional to variations in data density. These variations may be obtained by estimating variations in edge length in a path between two units on the SOM graph.

Given two neighboring SOM units, the edge length E is:

$$E(w_a, w_b) = \|w_a - w_b\|$$

The cost of a path P with *e* edges is calculated with:

$$C(a,b) = \sum_{j=1}^{j=e-2} \left| E(w_j, w_{j+1}) - E(w_{j+1}, w_{j+2}) \right|$$

The membership is calculated using:

$$u_{ab} = \frac{1}{1 + C(a,b)}$$

Areas with constant data density will have low costs and high memberships. Areas with varying data density will induce high costs, and thus low memberships.

Application of the method

In this work we used data at the ED level, referring to the metropolitan area of Lisbon. Training a SOM with this data leads to the U-Matrix of Fig 1. From the analysis of this figure, three prototype units were selected based on two principles: the selection of clearly differentiated clusters, *i.e.*, naturally distinct groups; and the selection of a small number of clusters, to keep to the essence of the method.

Prototypes 1 and 2 were chosen in the two distinct zones of outliers (dark areas in the U-Matrix) so as to represent extreme situations. Prototype 3 was chosen in a light area of the U-Matrix, indicating low variability in unit distance. A large number of units will have high membership to prototype 3, which will desirably represent the “average status” of EDs.

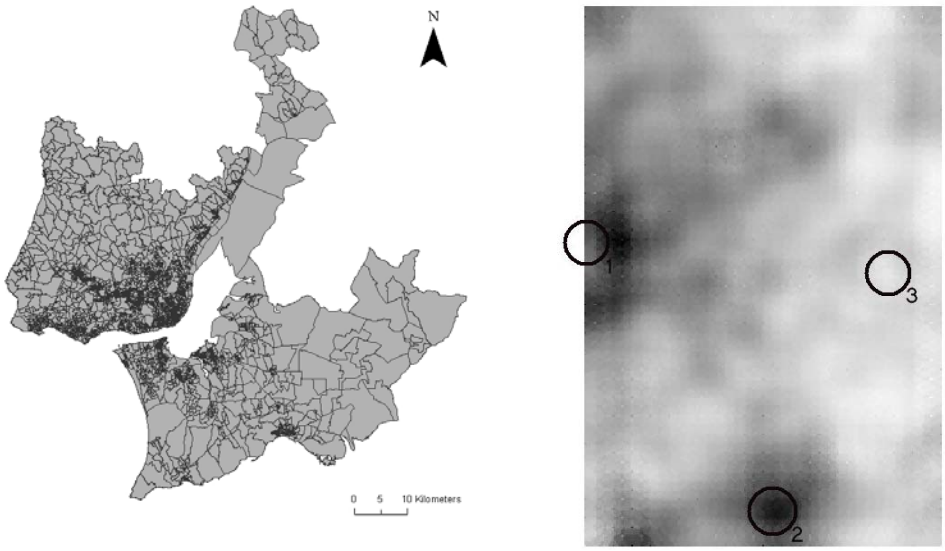

Fig 1. Left: the EDs of the metropolitan area of Lisbon. Right: U-matrix with prototype selection.

After selecting the prototypes, the membership of each SOM unit to each prototype unit is computed, which leads to a membership map like in Fig 2. Each original data point is given the same membership as its closest unit, allowing the representation of the EDs in Fig 2.

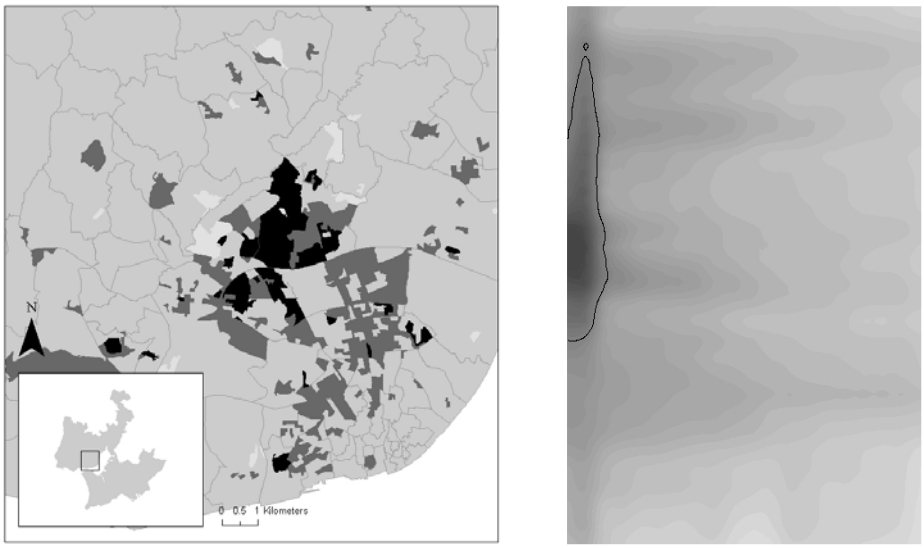

Fig 2. Fuzzy memberships to prototype 1. Left: EDs of the city of Lisbon. Right: membership map obtained from the U-Matrix

If at some point a crisp classification is needed, one may take each prototype and associate it to the cluster to which it has highest membership value, producing the crisp clustering of Fig 3.

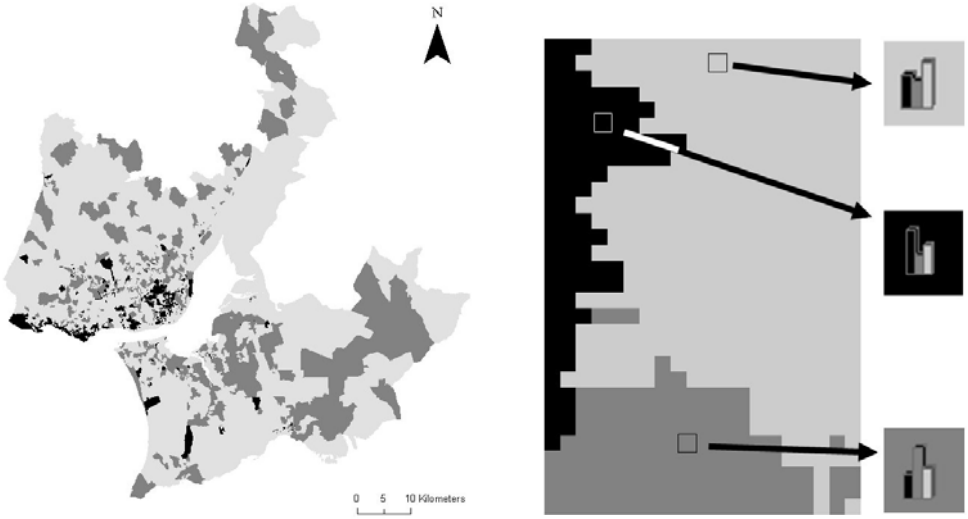

Fig 3. Crisp memberships to the three prototypes. Left: the metropolitan area of Lisbon. Right: crisp membership map

Conclusions

The new classification method proved to be a valid option in dealing with the major problems of geodemographics. The clusters may be set by visual inspection of the U-Matrix, which allows the selection of relevant prototypes. The proposed methodology is based on the output of a SOM, which is known to be robust to outliers and non-linear dependencies between variables. The fuzzy membership computation based on the U-Matrix introduces a new way of dealing with the fuzziness of the classification.

Tests with data from the Metropolitan Area of Lisbon have shown that the proposed methodology is adequate in the identification of its three main clusters. Cluster 1 represents affluent areas of Lisbon. Cluster 2 represents deprived areas. Cluster 3 represents the average situation of EDs, with a fairly balanced set of characteristics.

References

Bacao, F., Lobo, V., & Painho, M. (2005). Self-organizing Maps as Substitutes for K-Means Clustering. In V. S. Sunderam, G. v. Albada, P. Sloot & J. J. Dongarra (Eds.), Lecture Notes in Computer Science, (Vol. 3516, pp. 476-483). Berlin Heidelberg: Springer-Verlag.

Birkin, M., & Clarke, G. (1998). GIS, geodemographics and spatial modeling in the UK financial service industry. Journal of Housing Research, 9, 87-111.

Cottrell, M., Fort, J. C., & Pages, G. (1998). Theoretrical Aspects of the SOM algorithm. Neurocomputing, Elsevier, 21, 119-138.

Feng, Z., & Flowerdew, R. (1998). Fuzzy geodemographics: a contribution from fuzzy clustering methods. In S. Carver (Ed.), Innovations in GIS 5 (pp. 119-127). London: Taylor & Francis.

Fort, J. C. (2005). SOM's Mathematics. Paper presented at the WSOM'05 - Workshop on SOM, Paris.

Harris, R. (1998). Considering (mis-) representation in geodemographics and lifestyles. Paper presented at the 3rd International Conference on GeoComputation.

Kohonen, T. (2001). Self-Organizing Maps (3rd ed.). Berlin-Heidelberg: Springer.

Openshaw, S., & Wymer, C. (1994). Classifying and regionalizing census data. In S. Openshaw (Ed.), Census Users Handbook (pp. 239-270). Cambridge, UK: Geo Information International.