Comparing SaTScan and *Seg*-DBSCAN methods in spatial phenomena

Silvestro Montrone
Department of Statistics, University of Bari, s.montrone@dss.uniba.it

Paola Perchinunno
Department of Statistics, University of Bari, p.perchinunno@dss.uniba.it

Samuela L'Abbate
Department of Statistics, University of Bari, samuela.labbate@dss.uniba.it

Cosimina Ligorio
Department of Statistics, University of Bari, c.ligorio@dss.uniba.it

Abstract: The aim of this paper is to group territorial units in areas of high intensity, using SaTScan and *Seg*-DBSCAN clustering methods to aggregate adjacent spatial units that are homogeneous with respect to the phenomenon being studied. SaTScan scans the region of interest with a moving window and compares a smoothing of the intensity inside and outside it so that units belonging to contiguous windows with similar intensity are aggregated into a cluster. On the other hand, *Seg*-DBSCAN, a new version of DBSCAN, limits the arbitrariness of the choice of input parameters and identifies clusters as dense regions in space. As an application we analyze geo-referenced data concerning housing problems in Bari and we propose a comparison between the two methods presented.

Keywords: clustering, SaTScan, DBSCAN, *Seg*-DBSCAN, housing problems.

1. Introduction

Our work is prompted by the need to identify territorial areas and/or population subgroups characterized by situations of hardship or strong social exclusion through a fuzzy approach that allows the definition of a measure of the degree of belonging to the disadvantaged group. Grouping methods for territorial units are employed for areas with high (or low) intensity of the phenomenon by using clustering methods that permit the aggregation of spatial units that are both contiguous and homogeneous with respect to the phenomenon under study. This work aims to compare two different clustering methods: the first based on the technique of SaTScan and the other based on the use of *Seg*-DBSCAN, a modified version of DBSCAN.

2. SaTScan method

SaTScan scans the region of interest with a moving window and compares a smoothing of the intensity inside and outside it: units belonging to contiguous windows with similar intensity are aggregated into a cluster [2].

The identification of clusters means, therefore, to determine an area in which a set of points contributes to maximizing the incidence of the phenomenon within the area and to minimizing the incidence outside the area. In practice, the technique involves placing a monitoring window at random on the area of observation and then calculating the value of an estimator both inside and outside the area before proceeding to the testing of hypotheses.

3. *Seg*-DBSCAN method

DBSCAN (Density Based Spatial Clustering of Application with Noise) was the first density-based spatial clustering method proposed [1]. The key idea is that to define a new cluster or extend an existing cluster, a neighborhood around a point of a given radius ε must contain at least a minimum number of points *MinPts*, i.e. the density in the neighborhood is determined by the choice of a distance function for two points *p* and *q*, denoted by *dist(p,q)*. The greatest advantages of DBSCAN are that it can follow the shape of the clusters and that it requires only one distance function and two input parameters [1]. Their choice is crucial because they determine whether a group is a cluster of points or a simple noise.
In order to limit the arbitrariness of the choice of a value to assign to ε, usually detected by a heuristic procedure, in this work we develop a new algorithm: *Segmented* DBSCAN (*Seg*-DBSCAN), a modified version of DBSCAN, in which the clusters are aggregated considering multiple levels of value of ε.
Therefore, to define levels of ε, a value of *MinPts* is fixed and we analyze the distribution of the maximum radius of the cores that are groups formed by *MinPts* points. Then, we build a histogram of this distribution and we choose ε where there are the histogram peaks that indicate a proximity of the cores of a cluster. As suggested in literature, we can fix the value of *MinPts* to 4, and a number of levels of ε equal to the number of the highest histogram peaks.
The final phase of the algorithm is to merge the clusters obtained. The merging of two clusters C_1 and C_2 characterized by different levels of density ε_1 and ε_2 is obtained if

$$d\left(C_1, C_2\right) < \max(\varepsilon_1; \varepsilon_2) \qquad (1).$$

With this new algorithm, parameter ε is no longer established *a priori*.

3. Distance function for application

The aim of our study is to identify the dense areas in terms of intensity compared to the considered index. For this purpose, instead of Euclidean distance a function was chosen that warps the geometric space so that points that are geographically close and have a high intensity become even closer, while points that are geographically close, but at least one of which has a low intensity, become more distant.
The function that links in these terms two points *A* and *B* of coordinates $A(x_A, y_A, w_A)$ and $A(x_B, y_B, w_B)$ respectively, with $0 < \{w_A, w_B\} < 1$, is a weighted distance that is obtained by dividing the Euclidean distance by a mean of order integer $t>0$:

$$d_{pesata}(A,B) = \frac{\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}}{\sqrt[t]{\left(\frac{w_A^{-t} + w_B^{-t}}{2}\right)^{-1}}} \qquad (2).$$

Observe that in this distance the triangle inequality does not hold, so it is a semimetric, but this restriction does not affect the definitions of density-reachability and density-connectivity necessary for DBSCAN algorithm [1].
With this function the distance increases in matching pairs of points with low intensity value, so that they are penalized in the formation of clusters. Empirically it was verified that the most appropriate value of t is 5.

3. Application

This work aims to identify the land areas characterized by situations of housing problems by defining typical indicators able to estimate the difficulty in small areas. The case study uses data from the last Population and Housing Census carried out by ISTAT in 2001. The indices were calculated for each section of the census of the city of Bari [3]:

- incidence of the number of dwellings occupied by rent-payers with respect to the total number of dwellings occupied by residents;
- index of overcrowding: the ratio between the total number of residents and size of dwellings occupied by residents;
- availability of functional services: landline telephone, the presence of heating systems and the availability of a designated residential parking space.

These indices may be synthesized by a fuzzy index obtained by "Total Fuzzy and Relative" (TFR) method [3]; we denominate this new index “disadvantaged housing index”. It is a measure of an individual’s degree of membership to a disadvantaged group and its range is between zero (if the individual does not definitely belong to this group) and one (if the individual definitely belongs to this group).
Using the SatScan method, we identify different clusters each composed by a different number of sections of the city of Bari.
The city of Bari presents various critical areas: the old town of San Nicola, the areas surrounding the city center, Madonnella, Libertà and Carrassi (the former characterized by the presence of public housing complexes such as the Duca degli Abruzzi). Less critical, though more widespread, is the situation in some suburban areas such as Carbonara and Ceglie.
The same data on housing problems were analyzed with the *Seg*-DBSCAN method by associating geographic coordinates to the disadvantaged index to obtain eight clusters. The critical areas thus obtained do not exactly coincide with those identified by the SatScan method: both methods identified the old town of San Nicola and the areas surrounding the city center - Madonnella, Libertà and Carrassi – as well as Carbonara and Ceglie; but the *Seg*-DBSCAN method also identified the districts of San Cataldo and San Paolo
We observe that SatScan identifies areas formed by contiguous spatial units in which a smoothing of the disadvantaged housing index is performed. This method is effective in

identifying areas of high or low intensity and therefore may be a useful indication of areas "at risk" to be monitored.
Like the SatScan method, *Seg*-DBSCAN identifies areas in which the spatial units meet a criterion of adjacency, but *Seg*-DBSCAN differs in excluding those areas where the phenomenon is absent. *Seg*-DBSCAN can exactly identify sections of the city with housing problems. In the case of the San Nicola district, the old town of Bari, the SaTScan method identifies the whole district (Figure 1a) while the *Seg*-DBSCAN method identifies the same area of hardship but also analyzes the area in more detail (Figure 1b). The method identifies the particular points with a greater presence of the phenomenon and excludes the points where the phenomenon is not present because of the restoration of historic buildings.

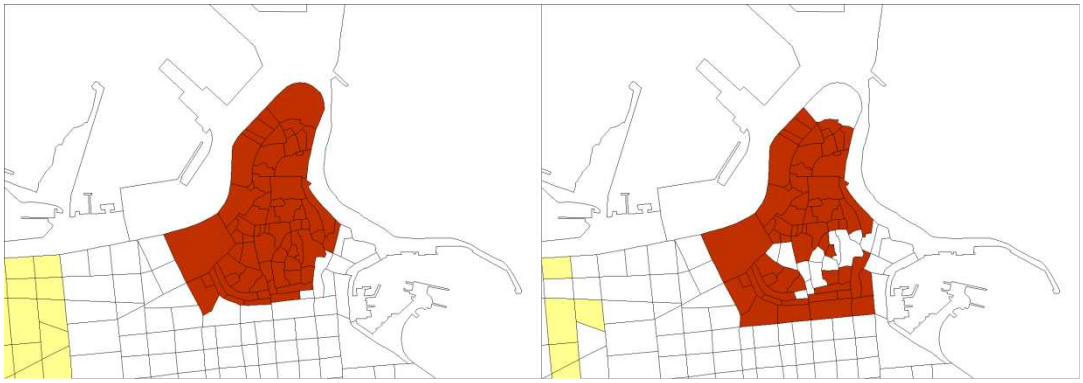

Figure 1a: SaTScan method **Figure 1b**: *Seg*-DBSCAN method

4. Conclusions and future advancements

The proposed methodologies identify areas where there is a high disadvantaged index. As we have noted above, a comparison of the two methods shows that the *Seg*-DBSCAN method is more accurate in identifying the spatial units in which there are housing problems. The future advancement of our work will be to seek a cluster validity index for spatial data, which takes into account the noise points, that is valid from a statistical point of view and that allows the accurate measurement of the *Seg*-DBSCAN method.

References

[1] Ester M., Kriegel H.-P., Sander J., Xu X. (1996) *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland.

[2] Kulldorff M.(1997) *A spatial scan statistic* Communications in Statistics Theory and Methods 26(6) 1481-1496.

[3] Montrone S., Perchinunno P., Di Giuro A., Rotondo F., Torre C.M. (2009) *Identification of "Hot Spots" of Social and Housing Difficulty in Urban Areas: Scan Statistics for Housing Market and Urban Planning Policies* in: *Geocomputation and Urban Planning*, Murgante B., Borruso G., Lapucci A. (Eds.), Springer, 57-78.