Bootstrap ClustGeo with spatial constraints

Bootstrap ClustGeo con vincoli spaziali

Veronica Distefano[1,2], Valentina Mameli[1], Fabio Della Marra[1,2]

Abstract The aim of this paper is to introduce a new statistical procedure for clustering spatial data when an high number of covariates is considered. In particular, this procedure is obtained by coupling the agglomerative hierarchical clustering method that ha been recently proposed for spatial data, referred as *ClustGeo* (*CG*) method , with the bootstrap technique. The proposed procedure, which we call *Bootstrap ClustGeo* (*BCG*), is developed and tested on a real dataset. The results that we achieve show that *BCG* outperforms *CG* in terms of accuracy of some cluster evaluation measures.

Abstract Il presente lavoro propone una nuova procedura di clustering per dati spaziali, che denoteremo *Bootstrap ClustGeo* (*BCG*). In particolare, questa nuova procedura coniuga il metodo di clustering agglomerativo gerarchico, recentemente proposto per dati spaziali sotto il nome di *ClustGeo* (*CG*), con il metodo bootstrap. I risultati ottenuti dimostrano una migliore performance dell'approccio *BCG* secondo un insieme di misure di valutazione di clustering.

Key words: Agglomerative hierarchical clustering, Bootstrap technique, *ClustGeo* method, Geographical data, Hamming distance, Spatial data.

1 Introduction

Addressing a study on sustainable development of geographical areas is becoming crucial to derive geopolitical policy. As noted by UN-GGIM (2012), "all of

[1] European Centre for Living Technology, Ca' Foscari University of Venice, Italy.
[2]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy.
e-mail: veronica.distefano@unive.it, e-mail: valentina.mameli@unive.it, e-mail: fabio.dellamarra@unive.it

the issues impacting sustainable development can be analyzed, mapped, discussed and modeled within a geographic context. Whether collecting and analyzing satellite images or developing geopolitical policy, geography can provide the integrative framework necessary for global collaboration and consensus decision-making" [11]. In this context, the geo-spatial clustering procedures represent an important area of research in data analysis, and a growing interest in clustering spatial data is emerging in several application fields. Indeed, the geo-spatial data and accurate clustering approaches in selecting constraints and parameters can provide better and more meaningful results.

In this study, we address the problem of clustering n spatial locations into K disjoint clusters when an high number of covariates is considered. Most approaches in the literature have been developed to derive clusters containing only contiguous locations. These procedures are based on the assumption that, within each cluster, there exists a connecting path for any couple of locations [9, 10, 5, 2]. This assumption involves then a strict-spacial constraint, i.e. locations characterized by social-economic variables with very similar values, but not close to each other in space, will be likely to be grouped into different clusters. Very recently, a non-strict constrained procedure has been developed, in which the condition of spatial closeness is relaxed [4]. In [6], the authors propose a hierarchical clustering method with non-strict spatial constraints, which is referred as *ClustGeo*. The method is based on two dissimilarity matrices: a matrix with dissimilarities derived from the "covariate-space"and a matrix with the dissimilarities derived from the "non-strict constraint space". In our work, we extend the aforementioned approach proposed in [6] by developing a procedure based on the generation of multiple bootstrap clustering partitions combined by using the Hamming distance. The novel procedure is called *Bootstrap ClustGeo*.

The paper is organized as follows. In Section, 2 we review the *ClustGeo* method and we then introduce the novel procedure *Bootstrap ClustGeo*. In Section 3, we evaluate this procedure on a real and known database which includes 303 French municipalities characterized by a set of 10 socio-economic covariates.

2 Methods

In classical cluster analysis, similar observations can be grouped into clusters. There exist different types of clustering algorithms which have been developed for different structure of data to be analyzed [8, 3]. In this paper, we select the agglomerative hierarchical clustering approach for analyzing geo-spatial data. The structure of an agglomerative hierarchical clustering approach can be summerised as follows. At the initialization, each cluster contains a single observation. Then, at each step of the agglomerative process, the two clusters with the smallest distance are merged into a new one. This procedure is iterated until a single cluster containing all the observations is obtained. The results of the agglomerative algorithm are usually represented with a tree or a dendrogram.

Formally, let $\{x_i = (x_{i1}, \ldots, x_{ip})\}_{i=1,\ldots,n}$ be the set of n observations (municipal-

ities in the dataset), each of which is described by p covariates. Let $\{w_i\}_{i=1,\ldots,n}$ be the weights associated to the i-th observation selected by the researcher. $\{w_i\}_{i=1,\ldots,n}$ can be proportional to the inverse of the total poluation of the municipality x_i. Consider $D_0 = (d_{0ij})_{i,j=1,\ldots,n}$ a $n \times n$ dissimilarity matrix associated with the n observations, where $d_{0ij} = d_0(x_i, x_j)$ is the dissimilarity measure between observations i and j, which could be Euclidean or non-Euclidean. In this paper we focus only on the non-Euclidean case. The matrix D_0 is usually referred as the dissimilarity of the "covariate-space". Consider also the $n \times n$ dissimilarity matrix $D_1 = \{(d_1(x_i, x_j))\}_{i,j=1,\ldots,n} = (d_{1ij})_{i,j=1,\ldots,n}$ containing spatial constraints between the observations, which is refereed as the dissimilarity of the "non-strict geo-spatial constraint space". In this research we consider the geographical distances as spatial constraints.

Let $\alpha \in [0,1]$ be a parameter which controls the importance of the spatial constraints in the clustering procedure. Suppose that the dataset, $X = (x_1, \ldots, x_n)^T$, is partitioned into K clusters C_k^α with $k = 1, \ldots, K$, which form the partition $\mathscr{P}_K^\alpha = (C_1^\alpha, \ldots, C_K^\alpha)$. In the agglomerative clustering process for spatial data, the distance between clusters could be determined by using the *ClustGeo* (*CG*) method as proposed in [6], which is based on the minimization of the pseudo within-cluster inertia of the partition $\mathscr{P}_K^\alpha$. The pseudo within-cluster inertia is defined as

$$W(\mathscr{P}_K^\alpha) = \sum_{k=1}^{K} I_\alpha(C_k^\alpha),$$

where

$$I_\alpha(C_k^\alpha) = (1-\alpha) \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i w_j}{2\mu_k^\alpha} d_{0ij}^2 + \alpha \sum_{i \in C_k^\alpha} \sum_{j \in C_k^\alpha} \frac{w_i w_j}{2\mu_k^\alpha} d_{1ij}^2,$$

with $\mu_k^\alpha = \sum_{i \in C_k^\alpha} w_i$. Starting from a given partition $\mathscr{P}_K^\alpha$, the *CG* method aggregates the two clusters with the smallest within-cluster inertia in order to obtain a new partition with $K-1$ clusters.

In order to analyze the database with an high number of covariates, in this paper we develop a novel procedure based on the *CG* method. This novel procedure is derived by generating multiple bootstrap clustering partitions with the *CG* method and combining the results by using the Hamming distance. We will refer to the new methodology as the *Bootstrap ClustGeo* (*BCG*) method. The main features of this novel methodology are described in the following.

At first, we take B bootstrap sample from the n data points x_i. Each bootstrap sample sample will be denoted by X_b, for $b = 1, \ldots, B$. For each bootstrap sample X_b, we will set the number of clusters K_b by drawing it from a discrete uniform distribution, i.e. $K_b \sim DUnif$. Then, we use the *CG* method to find a clustering partition $\mathscr{P}_{K_b}^\alpha$ of size K_b for each dataset X_b. At the end of the B bootstrap replications, we construct the incidence matrix

$$\mathscr{I}^{\alpha} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1B} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nB} \end{bmatrix},$$

where r_{ib} represents the index of the cluster in the partition $\mathscr{P}^{\alpha}_{K_b}$ to which x_i belongs. At last, we derive a new dissimilarity matrix

$$D^{\alpha}_{B} = (d^{\alpha}_{B}(x_i, x_j))_{i,j=1,\dots,n},$$

in which

$$d^{\alpha}_{B}(x_i, x_j) = \frac{1}{B} \sum_{b=1}^{B} \delta(r_{ib}, r_{jb}),$$

where δ is the Hamming distance. The new dissimilarity matrix D^{α}_{B} will be used to find a new clustering partition by exploiting the *CG* method.

3 Results

In this section, we compare the performances of *CG* and *BCG* methods on a real and known database. We perform a geo-spatial clustering analysis with geographical constraints in the administrative region of the Nouvelle Aquitanie, which is located in the southwest of France. The available dataset consists of n = 303 municipalities and $p = 10$ indicators, which includes social and economic indicators, as shown in Table 1. The data source of the indicators is the INSEE (*National Institute of Statistics and Economic Studies*).

The number of clusters has been estimated by visual inspection of the clustering tree generated by the *CG* algorithm when only D_0 is taken into account. The optimal value of the parameter α has been chosen by using the criterion proposed in [6]. From Table 1, we can see that the means of the vast majority of the variables within clusters 1 and 2 do not show significant differences across the two clustering methods. We can notice more relevant differences in the remaining clusters (3, 4, and 5). For example, the mean of the variable "Ratio of the agricultural area"assumes a small value (2.90) in cluster 5 by using the *CG* method with respect to the mean of the same variable (35.16) in cluster 3 with the *BCG* method.

From the results presented in Fig. 1, we can note that the clusters obtained by using the *BCG* are spatially more compact than those obtained by the *CG* method. Indeed, the municipalities located in cluster 2 for the *CG* method are grouped into a greater number of clusters than by the *BCG* method. Neither of the two methods requires any strict-contiguity assumption, as shown in Fig 1. Hence, cluster 1 contains municipalities which are not connected in space.

To assess the quality of the two clustering partitions obtained with the *CG* and the *BCG* algorithms, we consider some known evaluation measures, such as the Con-

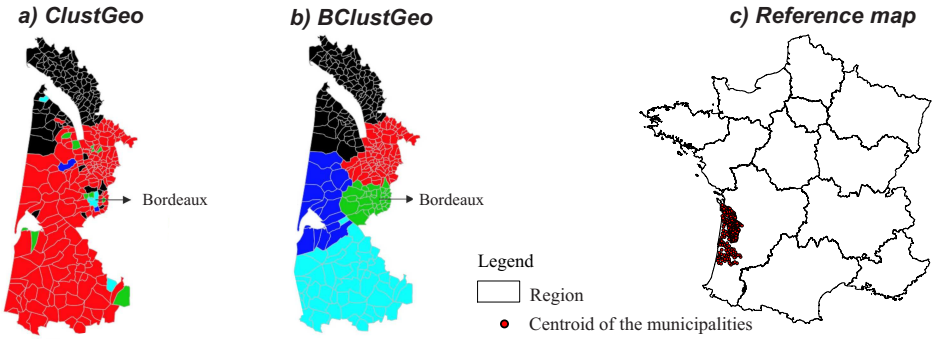
Indicators	Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
(S) Employment rate	*CG*	27.92 (11.09)	25.45 (13.41)	24.94 (12.48)	32.04 (0.42)	61.35 (20.98)
	BCG	28.34 (11.23)	24.28 (11.09)	24.94 (12.78)	31.70 (16.71)	61.35 (14.97)
(S) Level of Education	*CG*	15.43 (3.23)	14.88 (2.91)	17.12 (1.91)	16.43 (1.82)	17.28 (2.45)
	BCG	14.91 (3.08)	15.13 (2.68)	17.12 (2.69)	16.49 (3.27)	17.2 (3.15)
(S) Ratio of apartment housing	*CG*	6.37 (7.81)	5.11 (6.22)	33.32 (13.61)	10.38 (20.07)	74.39 (11.21)
	BCG	5.42 (9.24)	5.59 (11.85)	33.32 (18.39)	10.35 (13.37)	74.39 (9.63)
(S) Ratio of the agricultural area	*CG*	63.53 (25.18)	51.90 (29.33)	12.77 (8.15)	21.44 (1.30)	4.99 (2.19)
	BCG	64.45 (24.32)	45.85 (27.17)	12.77 (32.26)	19.60 (29.13)	4.99 (21.17)
(S) Average density of the population	*CG*	132.82 (182.43)	102.67 (98.32)	1480.56 (285.86)	195.83 (153.02)	4995.70 (330.07)
	BCG	109.15 (510.10)	110.97 (414.29)	1480.56 (1223.27)	201.23 (381.15)	4995.70 (662.32)
(E) Share of workplaces on business, transportations and in financial activities	*CG*	28.65 (19.83)	26.33 (19.99)	48.96 (8.20)	30.25 (10.32)	48.50 (6.31)
	BCG	24.48 (19.44)	29.09 (20.93)	48.96 (22.05)	30.74 (18.80)	48.50 (18.51)
(E) Share of workplaces on public administration, education, health and social action	*CG*	31.08 (19.06)	27.19 (16.91)	27.14 (10.53)	30.35 (8.49)	45.60 (10.39)
	BCG	29.28 (18.78)	27.14 (14.91)	27.14 (13.81)	30.47 (16.13)	45.60 (21.06)
(E) Share of employees on the total number of jobs	*CG*	25.13 (10.51)	25.11 (12.21)	26.57 (5.91)	25.85 (0.00)	29.80 (1.54)
	BCG	25.13 (10.64)	25.11 (11.82)	26.57 (9.11)	25.85 (9.25)	29.80 (13.97)
(E) Share of workers on the total number of jobs	*CG*	25.84 (13.25)	29.45 (13.98)	21.54 (5.11)	28.52 (1.41)	12.00 (4.32)
	BCG	27.00 (13.15)	30.01 (13.15)	21.54 (11.97)	28.47 (7.50)	12.00 (15.94)
(E) Share of owners in the main residences	*CG*	76.30 (8.30)	74.64 (9.01)	54.95 (6.29)	68.75 (9.19)	32.00 (11.59)
	BCG	76.68 (8.40)	73.45 (11.27)	54.95 (13.73)	68.57 (9.61)	32.00 (9.43)

Table 1: Comparison of the CG and the BCG methods in derived clusters for Economic (E) and Social (S) indicators. In each cluster we present the mean value and the standard deviation within brackets.

nectivity, the average Silhouette, and the Dunn indices [7]. The Connectivity represents the strength of connectedness of the clusters, lies in the range between 0 and infinity and should be minimized. Both the Silhouette and the Dunn indices measure the quality of the compactness and the separation of the the clusters and should be maximized. The Silhouette value lies in the interval $[-1,1]$, and the Dunn index assumes values between 0 and infinity. As we can see from Table 2, the *BCG* algorithm outperforms the *CG* algorithm for all the indices. These results show that the novel methodology proposed is capable of exploiting the spatial constraints to achieve better clustering accuracy in comparison with the *CG*. The use of the bootstrap sampling and the Hamming distance for categorical variables permit to accurately obtain information for defining the Clustering structure. The results show the BCG to be more accurate then the CG.

		Indices		
		Silhouette	Dunn	Connectivity
Methods	*CG*	0.07	0	117
	BCG	0.79	0.64	0

Table 2: Evaluation measures to validate the clustering methods.



d) Comparison of the CG and BCG methods

	Nr.municipalities	
	ClustGeo	*BClustGeo*
Cluster 1	108	112
Cluster 2	178	90
Cluster 3	11	32
Cluster 4	2	19
Cluster 5	4	50

Fig. 1: Study area and maps generated by the *CG* and the *BCG* algorithms.

References

1. Amiri, S., Clarke, B.S., Clarke, J.L.: Clustering categorical data via ensembling dissimilarity matrices. J. Comput. Graph. Statist. 1–14 (2017).
2. Becue-Bertaut, M., Alvarez-Esteban, R., Sanchez-Espigares, J.A., Xplortext: Statistical Analysis of Textual Data R package. https://cran.r-project.org/package=Xplortext. R-package version 1.0 (2017).
3. Benassi, F., Bocci, C. and Petrucci, A.. Spatial data mining for clustering: an application to the Florentine Metropolitan Area using RedCap. Classification and Data Mining, pp. 157-164. Springer, Berlin, Heidelberg (2013)
4. Bourgault, G., Marcotte, D., Legendre, P.: The Multivariate (co) Variogram as a Spatial Weighting Function in Classification Methods. Mathematical Geology 24(5): 463–478 (1992).
5. Carvalho, A. X. Y., Albuquerque, P. H. M., de Almeida Junior, G. R., Guimaraes, R. D.: Spatial hierarchical clustering. Revista Brasileira de Biometria, 27(3), 411–442 (2009).
6. Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J.: ClustGeo: an R package for hierarchical clustering with spatial constraints. Computational Statistics, 1-24 (2018).
7. Brock, G., Pihur, V., Datta, S., Datta, S.: clValid: An R Package for Cluster Validation. Journal of Statistical Software 25: 1–22 (2008)
8. Everitt, B., Landau, S., Leese, M., Stahl, D.:Cluster analysis. 5th edn, Wiley, Chichester (2011).
9. Lance, G.N., Williams, W.T.: A General Theory of Classicatory Sorting Strategies 1. Hierarchical Systems. The Computer Journal 9: 373–380 (1967).
10. Murtagh, F.: Multidimensional clustering algorithms. Compstat Lectures, Vienna: Physika, Verlag (1985).
11. UN-GGIM. 2012. Monitoring Sustainable Development: Contribution of Geospatial Information to the Rio Processes. New York: United Nations. Accessed January 17, (2016).