Clustering of spatio-temporal data based on marked variograms

Clustering di dati spazio-temporali basato su marked variograms

Antonio Balzanella and Rosanna Verde

Abstract This paper deals with the clustering of data generated by spatio-temporal point processes. The interest on this topic is motivated by the recent availability of spatio-temporally indexed data in several applicative fields like seismology, climatology, economics, social sciences. The data we analyse is a collection of instantaneous events, each occurring at a given spatial location. We introduce a strategy which finds a partition of the individuals into homogeneous clusters considering the space-time interactions. We transform the spatio-temporal point process into two marked point processes, considering the times as marks of the spatial point process and the locations as marks of the times. This allows to use the marked variograms to describe the second-order characteristics of the individuals, in time and space. We propose a k-means like algorithm which uses the marked variograms as cluster representative and performs the allocation to clusters evaluating the contribution of each individual to the definition of the marked variograms. This allows to get clusters of individuals which are homogeneous in terms of space-time interactions.
Abstract *Il presente articolo è incentrato sulle tecniche di clustering per dati generati da processi di punto spazio-tempo. L'interesse sulla tematica è dovuto alla crescente disponibilità di dati spazio-tempo in ambiti applicativi quali la sismologia, la climatologia, le scienze economiche, le scienze sociali. La tipologia di dati che si propone di analizzare è una collezione di eventi istantanei verificatisi, ciascuno, in una specifica locazione spaziale. In tale contesto, si propone una strategia di clustering finalizzata all'individuazione di una partizione degli individui in cluster omogenei, considerando le interazioni spazio-tempo. Si propone di trasformare il processo di punto spazio-tempo in due processi di punto marcati nei quali si considera il tempo, come marcatore di un processo di punto spaziale e lo spazio, come marcatore di un processo di punto temporale. Ciò consente di utilizzare i var-*

Antonio Balzanella
Università della Campania L. Vanvitelli, e-mail: antonio.balzanella@unicampania.it

Rosnna Verde
Università della Campania L. Vanvitelli e-mail: rosanna.verde@unicampania.it

iogrammi marcati per descrivere le caratteristiche di secondo ordine degli individui nel tempo e nello spazio. Nella strategia proposta, i variogrammi marcati vengono utilizzati come centroidi di un algoritmo di tipo k-means che effettua l'allocazione ai cluster sulla base del contributo di ciascun individuo alla definizione della struttura di variabilità spazio-temporale sintetizzata dai variogrammi marcati. Tale approccio consente di ottenere cluster di individui che sono omogenei in termini di interazioni spazio-tempo.

Key words: clustering, spatio-temporal point process, mark variogram

1 Introduction

In this paper we deal with the statistical analysis of a collection of data generated by a spatio-temporal point process. Single observations are instantaneous events, each occurring at a given spatial location with a given associated time stamp. Typical applications are the analysis of seismic events, the monitoring of crimes or diseases.

Recent proposals in [2][4] provide, a review of the statistics for analysing such kind of data and new methods for exploring the spatio-temporal data structure by means of suitable second-order statistics.

We focus on clustering the data points into homogeneous groups keeping into consideration their spatio-temporal interactions.

Consistently with [4], we transform the spatio-temporal point process into two marked point processes. We can consider the times as marks of the spatial point process of point locations and the locations as marks of the times.

For evaluating the second-order characteristics of the two marked point processes, we use the mark variograms. The latter, provide a measure of the data interaction over the whole geographic area, for the considered time window.

In this paper, we propose a decomposition of the mark variograms into a set of local measures of spatio-temporal interaction (LMSTI), which allow to evaluate the contribution of each data point to the definition of the mark variogram.

Since we use two mark variograms for evaluating the second order characteristics of the spatio-temporal point process, we will associate each data point to two LMSTI, the first one uses the time as mark, the second one uses the location as mark.

Our idea, is to describe each data point through the two mark LMSTI and to perform a k-means like algorithm on this new data description. This allows to get clusters of data points which interact similarly with the other data points, accounting both for the space and time dimension.

2 Main notations and definitions

We consider a spatio-temporal point process $X = \{(s_1,t_1),\ldots,(s_i,t_i),\ldots,(s_n,t_n)\}$, where $s_i \in D \subseteq \Re^2$ is a spatial location and $t_i \in T \subseteq \Re^+$ the corresponding time. We assume that the point process is orderly, that is, coincident points cannot occur.

Similarly to [4], the random process X can be transformed into two mark point processes X_t and X_s using, respectively, the time and the locations as marks.

This allows to use the classic mark variogram [3] for measuring the interactions in the two processes. In particular, if we consider the time, as mark, the variogram γ_{sp} is:

$$\gamma_{sp}(h) = \frac{1}{2}\mathbf{E}\left[(t_i - t_j)^2 \;\middle|\; t_i, t_j \in T\right] \tag{1}$$

where $h = \left\|s_i - s_j\right\|$.

Similarly, if we consider the spatial location as mark, the variogram γ_{ti} is:

$$\gamma_{ti}(v) = \frac{1}{2}\mathbf{E}\left[(s_i - s_j)^2 \;\middle|\; s_i, s_j \in D\right] \tag{2}$$

where $v = \left\|s_i - s_j\right\|$.

Similarly to the classic variogram used in geostatistics ([1]), the mark variogram gives information on the correlations in the marked point process.

In order to estimate $\gamma_{sp}(h)$ we can use the expression in [4]:

$$\hat{\gamma}_{sp}(h) = \frac{\sum_{s_i,s_j \in D} \frac{1}{2}(t_i - t_j)^2 k_e(\left\|x_i - x_j\right\| - h)}{\sum_{s_i,s_j \in D} k_e(\left\|x_i - x_j\right\| - h)} \tag{3}$$

where k_e is is a one-dimensional kernel function with bandwidth e.

Similarly, $\gamma_{ti}(v)$ can be estimated by:

$$\hat{\gamma}_{ti}(v) = \frac{\sum_{t_i,t_j \in T} \frac{1}{2}(s_i - s_j)^2 k_g(\left\|t_i - t_j\right\| - v)}{\sum_{t_i,t_j \in T} k_g(\left\|t_i - t_j\right\| - v)} \tag{4}$$

where k_g is a one-dimensional kernel function with bandwidth g.

In order to introduce our clustering algorithm, we associate to each data point two functions which measure the interaction of the data point (s_i,t_i) with the other data points. The two functions Δ^i_{sp} and Δ^i_{ti} are obtained by decomposing the mark variograms γ_{sp} and γ_{ti}:

$$\Delta^i_{sp}(h) = \sum_{s_j \in D} \frac{1}{2}(t_i - t_j)^2 k_e(\left\|x_i - x_j\right\| - h) \tag{5}$$

$$\Delta^i_{ti}(v) = \sum_{t_j \in T} \frac{1}{2}(s_i - s_j)^2 k_g(\left\|t_i - t_j\right\| - v) \tag{6}$$

where $h = \left\|s_i - s_j\right\|$ and $v = \left\|s_i - s_j\right\|$.

3 Clustering of local measures of spatio-temporal interaction

By means of Δ^i_{sp} and Δ^i_{ti} we get a new description of the observed data points which we use in a k-means like algorithm for obtaining a partitioning P of the data points (s_i, t_i) into K clusters C_k (with $k = 1, \ldots, K$).

In order to reach our aim, we propose to minimize the following objective function:

$$J(P,L) = \sum_{k=1}^{K} \sum_{(s_i,t_i)\in C_k} d^2(\Delta^i_{sp}(h); \overline{\Delta^k_{sp}}(h)) + d^2(\Delta^i_{ti}(v); \overline{\Delta^k_{ti}}(v)) \tag{7}$$

where:

$\overline{\Delta^k_{sp}}(h)$ and $\overline{\Delta^k_{ti}}(v)$ are the prototypes of the cluster C_k;
L is the matrix of prototypes;
$d^2(.)$ is the squared euclidean distance.

The minimization of the objective function is performed by an iterative algorithm which starts from an initial random partitioning of the data points and, then, alternates a representation step and an allocation step until a stable value of $J(P,L)$ is reached.

In the representation step, the prototype $\overline{\Delta^k_{sp}}$, $\overline{\Delta^k_{ti}}$ of each cluster are computed by:

$$\overline{\Delta^k_{sp}} = \frac{\sum_{(s_i,t_i)\in C_k} \Delta^i_{sp}}{|C_k|} \tag{8}$$

$$\overline{\Delta^k_{ti}} = \frac{\sum_{(s_i,t_i)\in C_k} \Delta^i_{ti}}{|C_k|} \tag{9}$$

that is, they are the average of the LMSTI functions allocated to a cluster and can be seen as a summary of the spatio-temporal interaction structure in each cluster.

In the allocation step, the data points (s_i, t_i) are allocated to the cluster C_k, if the squared Euclidean distance between LMSTI functions Δ^i_{sp} and Δ^i_{ti} and the cluster prototype $\overline{\Delta^k_{sp}}$, $\overline{\Delta^k_{ti}}$ is minimum, through the following rule:

$$\begin{aligned} d^2\left(\Delta^i_{sp}(h); \overline{\Delta^k_{sp}}(h)\right) + d^2\left(\Delta^i_{ti}(v); \overline{\Delta^k_{ti}}(v)\right) < \\ d^2\left(\Delta^i_{sp}(h); \overline{\Delta^{k'}_{sp}}(h)\right) + d^2\left(\Delta^i_{ti}(v); \overline{\Delta^{k'}_{ti}}(v)\right) \quad \forall k \neq k' \end{aligned} \tag{10}$$

Since the centroids of each cluster are the average of the allocated items, similarly to the classic k-means, the optimized criterion decreases with the iterations until it reaches a stable value.

4 Preliminary results on real data

In this section, we introduce some preliminary results on real data. The dataset collects earthquakes in Japan from 1926 to 2005. The dataset contains 948 epicenters of earthquakes with a magnitude higher than 4.5. The data was obtained from `http://users.jyu.fi/~penttine/ppstatistics/data/Earthquakes_Fig_6_21.txt`. We consider D as a planar rectangle of side lengths $a = 543km$ and $b = 556km$, and a time interval T of length 12000 days.

We show in figure 3 the results of the partitioning into $K = 4$ clusters of the dataset of seismic events.

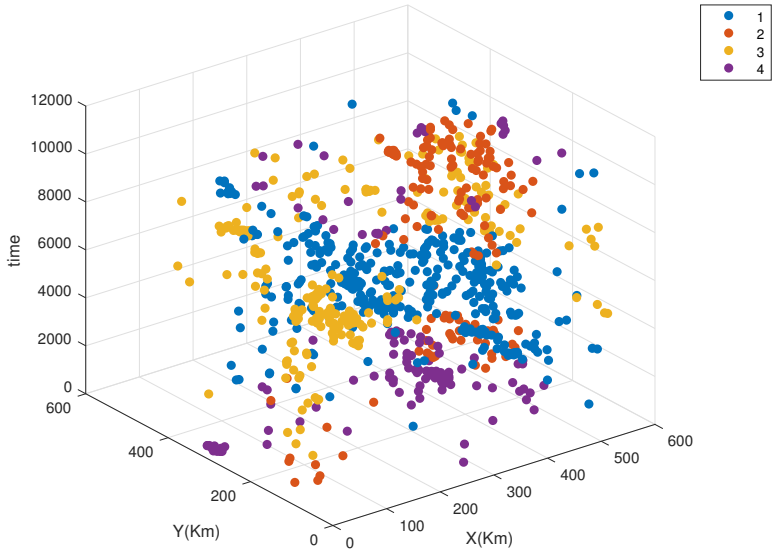

Fig. 1 Partitioning of the seismic events into $K = 4$ clusters.

Still, we show the LMSTI functions and the their average (cluster prototype) for each cluster and for the space and time domain in fig.

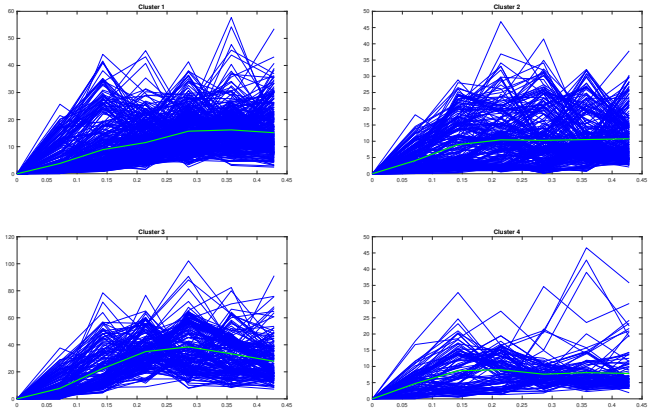

Fig. 2 LMSTI functions and cluster centroid in Space domain

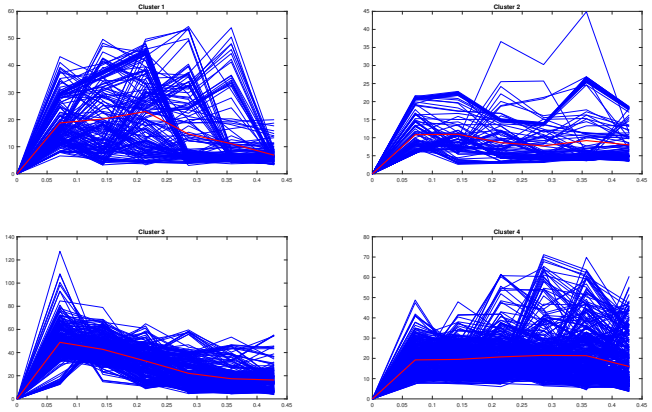

Fig. 3 LMSTI functions and cluster centroid in Time domain

References

1. Cressie N. Statistics for Spatial Data. John Wiley & Sons, New York (1993)
2. González J. A., Rodríguez-Cortés F.J., Cronie O., Mateu J.: Spatio-temporal point process statistics: A review. Spatial Statistics, **18**, Part B, 505–544 (2016) doi: https://doi.org/10.1016/j.spasta.2016.10.002.
3. Illian J., Penttinen A., Stoyan H., Stoyan D. In: Statistical Analysis and Modelling of Spatial Point Patterns, John Wiley & Sons, Chichester (2008)
4. Stoyan D., Rodríguez-Cortés F. J., Mateu J., Gille W.:Mark variograms for spatio-temporal point processes. Spatial Statistics, **20**, 125–147 (2017) doi: https://doi.org/10.1016/j.spasta.2017.02.006.