Classification of spatio-temporal point pattern in the presence of clutter using K-th nearest neighbour distances

Classificazione dei processi puntuali spazio-temporali basata sulla distanza dal K-mo vicino più vicino

Siino Marianna, Francisco J. Rodríguez-Cortés, Jorge Mateu, Giada Adelfio

Abstract In a point process spatio-temporal framework, we consider the problem of features detection in the presence of clutters. We extend the methodology of Byers and Raftery (1998) to the spatio-temporal context by considering the properties of the K-th nearest-neighbour distances. We make use of the spatio-temporal distance based on the Euclidean norm where the temporal term is properly weighted. We show the form of the probability distributions of such K-th nearest-neighbour distance. A mixture distribution, whose parameters are estimated with an EM algorithm, is used to classify points into clutters or features. We assess the performance of the proposed approach with a simulation study, together with an application to earthquakes.

Abstract *Nell'ambito dei processi puntuali spazio-temporali, abbiamo affrontato il problema relativo all'identificazione di aggregazione di punti in presenza di di eventi di fondo. Abbiamo esteso la metodologia proposta da Byers and Raftery (1998) nel contesto spazio-temporale considerando le proprietà (quali la distribuzione di probabilità) della distanza dal K-mo vicino più vicino. In particolare, abbiamo fatto uso della distanza Euclidea (dove la parte temporale è opportunamente pesata con un fattore di scala) mostrando la distribuzione di probabilità di questa distanza. Un approccio basato sulla misture di distribuzioni è stato utilizzato per classificare i punti nelle due rispettive categorie, utilizzando il metodo di risoluzione EM. Con-*

Siimo Marianna
Istituto Nazionale di Geofisica e Vulcanologia, Centro Nazionale Terremoti, Rome, Italy, e-mail: marianna.siino@ingv.it

Francisco J. Rodríguez-Cortés
Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia e-mail: frrodriguezc@unal.edu.co

Jorge Mateu
Departament of Mathematics, University Jaume I, Castellón, Spain e-mail: mateu@uji.es

Giada Adelfio
Dipartimento di Scienze Economiche, Aziendali e Statistiche,
Università degli Studi di Palermo, Palermo, Italy e-mail: giada.adelfio@unipa.it

siderando diversi scenari, abbiamo verificato il comportamento del metodo proposto. Inoltre, abbiamo considerato un'applicazione in ambito sismico per la classificazione dei terremoti in eventi di fondo e indotti.

Key words: Clutter; Earthquakes; EM algorithm; Features; Mixtures; Nearest-neighbour distances; Spatio-temporal point patterns.

1 Introduction

One of the most important research fields of spatio-temporal data-mining is the identification of features (clusters) of events. In particular, features are defined as subgroups of events in constrained spatio-temporal volumes with a higher density than other events outside the spatio-temporal windows (called background, noise or clutter events). The identification of such spatio-temporal features may yield insight for many applications. In practice, many geographical phenomena (e.g. earthquakes, disease cases, crime data, forest fires) are modelled as spatio-temporal events, and the detection of features is used to study the evolution of the phenomena, to reveal space or time anomalies and spatio-temporal hotspots. However, the detection of features is a challenge problem for the complexity caused by the time-space coupling and the noise interference.

For point processes, the corresponding problem has been widely addressed from a spatial point of view (Allard and Fraley, 1997; Byers and Raftery, 1998; Dasgupta and Raftery, 1998; Illian et al, 2008). For instance, Byers and Raftery (1998) estimated and removed the clutter without making any assumptions about the shape or number of features. Their method of detection is based on the distance to the K-th nearest-neighbour of a point in a spatial process. The observed distances are modelled as a mixture distribution of distances coming from clutter and feature points, the parameters of which are estimated by an EM algorithm.

In this paper, we use spatio-temporal distances obtained as a weighted version (with respect to the temporal component) of the Euclidean distance (Demattei and Cucala, 2010). Moreover, it is shown that the distribution of the K-th nearest-neighbour based on the previous distance follows an inverse Gamma distribution under the homogeneous Poisson assumption. The spatio-temporal K-th nearest-neighbour distance is analysed through a mixture model formulation of the corresponding distance densities coming from the clutter and feature events. The corresponding parameters associated to the two density distributions in the mixture model formulation are estimated using an expectation-maximisation (EM) algorithm.

A simulation study is carried out to assess the performance of the proposed classification method. Our method is also compared with the results obtained with the spatial methodology of Byers and Raftery (1998) in terms of sensitivity, specificity and accuracy. Finally, we present a seismic application, identifying noise and feature events in the seismic sequences occurred in California (near the Landers town, in 1992).

2 Methodology

We consider a spatio-temporal point process with no multiple points as a random countable subset X of $\mathbb{R}^{d-1}\times\mathbb{R}$, where a point $(\mathbf{u},s)\in X$ corresponds to an event at $\mathbf{u}\in\mathbb{R}^{d-1}$ occurring at time $s\in\mathbb{R}$. We observe n events $\{(\mathbf{u}_i,s_i)\}_{i=1}^{n}$ of distinct points of X within a bounded spatio-temporal region $W\times T\subset\mathbb{R}^{d-1}\times\mathbb{R}$, with volume $|W|>0$, and with length $|T|>0$ where $n\geq 0$ is not fixed in advance. We assume that the point process X is stationary and isotropic. Let $(\mathbf{u},s)=(u_1,u_2,\ldots,u_{d-1},s)$, and $(\mathbf{v},l)=(v_1,v_2,\ldots,v_{d-1},l)$ be points of a spatio-temporal homogeneous Poisson process in $W\times T\subset\mathbb{R}^{d-1}\times\mathbb{R}$. Following Dematteiand Cucala (2010) and given $d=2$, we consider the spatio-temporal Euclidean distance given by

$$D^{ST_E}\left((\mathbf{u},s),(\mathbf{v},l))\right)=\sqrt{(u_1-v_1)^2+(u_2-v_2)^2+\rho^2|s-l|^2}, \tag{1}$$

that is a mixture of a spatial and temporal distances where ρ is a scaling coefficient for the temporal part. The close form of the probability distribution function for the K-th nearest-neighbour Euclidean distance (Siino et al, 2019), under a homogeneous spatio-temporal Poisson process is given by

$$D_K^{ST_E}\sim\Gamma^{1/d}\left(K,\frac{\pi^{d/2}\lambda}{\Gamma\left(d/2+1\right)\rho}\right). \tag{2}$$

The maximum likelihood estimate (MLE) of the rate of the process is given by $\hat{\lambda}=K/(\alpha_d\sum_{i=1}^{N}\gamma_i^n)$ where the γ_i are the observations of $D_K^{ST_E}$, N is the sample size and $\alpha_d=(\pi^{d/2})/(\Gamma\left(d/2+1\right)\rho)$.

As in Byers and Raftery (1998) and Mateu et al (2007), we assume to have two types of processes to be classified, and model the K-th nearest-neighbour distances through a mixture of the corresponding K-th nearest-neighbour distances coming from the clutter and feature, which are two superimposed spatio-temporal Possion processes. We then suppose that the distribution of the $D_K^{ST_E}$ is roughly a mixture of distributions

$$D_K^{ST_E}\sim q\Gamma^{1/d}\left(K,\alpha_d\lambda_1\right)+(1-q)\,\Gamma^{1/d}\left(K,\alpha_d\lambda_2\right), \tag{3}$$

where λ_1 and λ_2 are the intensities of the two homogeneous spatio-temporal Poisson point processes (clutter and feature) and q is the constant which characterises the postulated distribution of the $D_K^{ST_E}$. The corresponding parameters associated to this mixture are estimated using an expectation-maximisation (EM) algorithm, where in the expectation step we use the close form provided by an inverse Gamma distribution.

3 Simulation study

A simulation study is carried out to assess the performance of the proposed methodology in terms of detection of features in a spatio-temporal setting. The spatio-temporal classification procedure proposed in this paper is compared with the method based on the spatial K-th nearest-neighbour distance in Byers and Raftery (1998), named $M_{spatial}$, as if we ignore time. The spatio-temporal window is set as

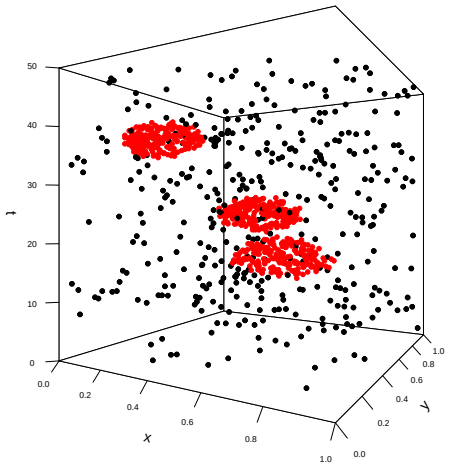


Fig. 1: Simulated scenarios with $n_c = 400$ clutter points from a homogeneous Poisson point process in a spatio-temporal window $[0,1]^2 \times [0,50]$. There are three ellipsoids ($n_{clusters} = 3$) with semi-axis $a = 0.2$, $b = 0.15$ and $c = 3.5$ with volume 0.43, each with $n_{fc} = 200$ points, and $n_f = 600$.

$W \times T = [0,1]^2 \times [0,50]$, where the time range is greater than the spatial one to have a different scale than the spatial window. In the spatio-temporal window, the clutters and features are simulated from two different processes. The clutter points are taken from a spatio-temporal homogeneous Poisson process with size $n_c = \{400, 1000\}$. There are three clusters ($n_{cluster} = 3$) and the number of feature points for each cluster is $n_{fc} = \{200, 400\}$, so the total number of feature points is $n_f = n_{fc} \times n_{cluster}$. Figure 1 shows a simulated point pattern for $n_c = 400$ clutter points and a number of feature points per cluster of $n_{fc} = 200$, with three ellipsoids.

We report results for $K = \{5, 10\}$ and $\rho = \{0.02, 0.5, 1\}$. Parameter ρ rescales the temporal distance, such that the weight of the temporal term changes accordingly. When $\rho = 1$, the methodology is equal to $M_{spatial}$ in three dimensions (Byers and Raftery, 1998). Instead, when $\rho = 0.02$, it corresponds to simulate the points in the unit cube. For each scenario, 100 point patterns are simulated as described above. The evaluation and comparison of the classification procedures with respect to the

different settings are done in terms of the true positive rate (TPR), the false positive rate (FPR) and the accuracy (Stehman, 1997). The results are shown in Table 1.

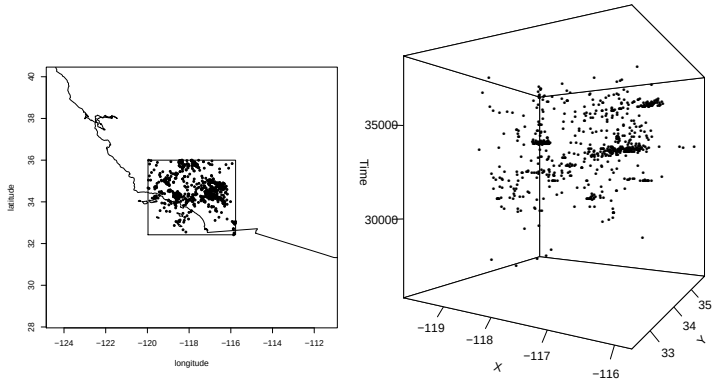
Our method always outperforms $M_{spatial}$, in terms of TPR, FPR and accuracy. Decreasing the value of ρ, so making closer the unit measurements of the space and time dimensions, the TPR, the specificity (1-FPR) and accuracy increase. In general, we observe that changing the value of K (from 5 to 10) the overall performance is comparable. Moreover, as expected, increasing the number of clutter points in $W \times T$ ($n_c = 1000$), the results are in general slightly worse than the cases with $n_c = 400$. For further simulation results and comparisons see Siino et al (2019).

Table 1: Results in terms of true positive rate (first row), false positive rate (second row) and accuracy (third row) in percentage over 100 simulated point patterns when the spatio-temporal feature points are three ellipsoids. The total number of clutter points is $n_c = \{400, 1000\}$. $n_{fc} = \{200, 400\}$ indicates the number of points for each feature, K is the K-th nearest-neighbour distance ($K = \{5, 10\}$) and $\rho = \{1, 0.5, 0.02\}$ is the weight for the temporal term in the spatio-temporal distance. D^{ST_E} refers to the Euclidean space-time distance. $M_{spatial}$ indicates the results obtained with the spatial method of Byers and Raftery (1998), neglecting time.

		$n_c = 400$				$n_c = 1000$			
		$K = 5$		$K = 10$		$K = 5$		$K = 10$	
n_f	ρ	D^{ST_E}	$M_{spatial}$	D^{ST_E}	$M_{spatial}$	D^{ST_E}	$M_{spatial}$	D^{ST_E}	$M_{spatial}$
200	1	97.72	96.24	96.16	97.80	97.51	92.87	96.50	95.26
		8.82	26.89	11.89	26.42	9.20	33.07	10.43	27.44
		95.11	86.99	92.94	88.12	93.32	76.66	92.17	81.07
	0.5	98.97	-	98.03	-	98.59	-	98.20	-
		6.64	-	8.33	-	6.76	-	7.59	-
		96.73	-	95.49	-	95.25	-	94.58	-
	0.02	99.83	-	99.96	-	99.54	-	99.88	-
		4.80	-	5.72	-	4.49	-	5.05	-
		97.98	-	97.69	-	97.02	-	96.80	-
400	1	99.08	97.75	98.31	98.78	98.79	95.64	98.09	97.04
		7.50	26.00	9.72	26.43	6.65	26.93	8.12	25.01
		97.43	91.81	96.30	92.48	96.32	85.38	95.27	87.02
	0.5	99.62	-	99.26	-	99.43	-	99.05	-
		5.81	-	7.26	-	5.37	-	6.39	-
		98.26	-	97.63	-	97.25	-	96.58	-
	0.02	99.96	-	100.00	-	99.83	-	99.97	-
		4.46	-	5.28	-	4.04	-	4.61	-
		98.86	-	98.68	-	98.07	-	97.89	-

4 Application on California earthquakes

In this section, the proposed method is applied to seismic data. Since an earthquake can be viewed as a spatio-temporal pattern, the identification of clustered earthquakes provides key information on seismic dynamics. Well-studied statistical models are based on the idea that the seismicity can be considered as the sum of "background" earthquakes (caused by tectonic loading) and "triggered" earthquakes (Ogata, 1988; Adelfio and Chiodi, 2015). To describe the seismicity of an area in space, time and magnitude domains, sometimes it is useful to study separately the features of *independent* events and *triggered* ones. At this regard, the proposed method based on the EM algorithm allows the identification of these two main components, such that the background seismicity is related to the long-term analysis, and the triggered one for sequence identification.

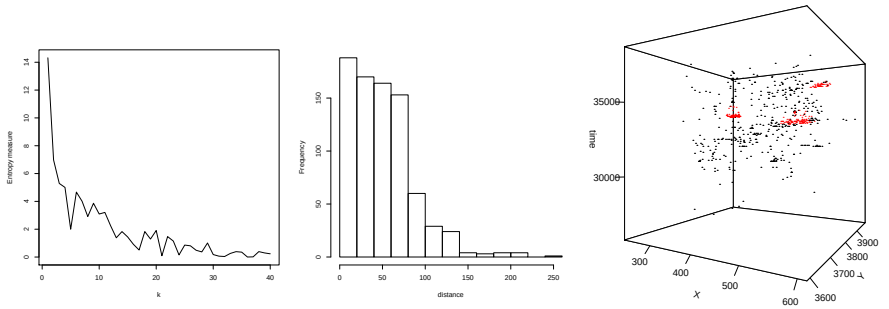


(a) (b)

Fig. 2: 2D and 3D-plots of the earthquake near Landers city, California.

Our goal is to detect spatio-temporal features (clusters) of earthquakes that occurred in California choosing a spatial window around high seismic areas. The subset of the earthquake catalogue in California refers to a study area located between -120–115 N and 32-36 E, near the Landers town, where in June 25^{th}, 1992 an event with magnitude 7.3 occurred, causing severe damage to the area directly surrounding the epicenter. A total number of 804 were observed with magnitude at least 3.5 from 1968 to 2012 (see Figure 2).

The value of ρ is set as the ratio between the maximum spatial distance over the maximum temporal distance observed between the events, so $\rho_{california} = 0.042$. The entropy measure ($S = \sum_{i=1}^{n} \delta_i log(\delta_i)$ where δ_i are the probabilities of being in the feature group) for each value of K is reported, Figure 3a. We selected the value of $K = 25$ since after this value the entropy measure can be considered constant. The histograms of the selected K-th nearest-neighbour distance based on the Euclidean



(a) (b) (c)

Fig. 3: (3a) Entropy measure changing the K-th value of the K-th nearest-neighbour using D^{ST_E}. (3b) Histogram of the D^{ST_E} distance to the 25-th nearest-neighbour. (3c) Detected feature and clutter points with $\rho_{california}$ in D^{ST_E}.

weighted distance is in Figure 3b. The results applying the EM classification procedure are reported in Figure 3c, where we can see that the feature points are clearly identified. The total number of clutter and feature points were $n_c = 477$ and $n_f = 327$ and three spatio-temporal features are identified.

5 Conclusions

In this paper, we present a classification method for identifying regions with higher point densities (features) in a spatio-temporal context extending the procedure of Byers and Raftery (1998) that is based on the spatial K-th nearest-neighbour distance. We use the weighted Euclidean spatio-temporal distance where the temporal term is scaled to account for the space-time coupling. Based on these results, a mixture model formulation for the K-th nearest-neighbour distance of clutter and feature points is considered to perform a binary classification using an iterative EM algorithm. With a simulation study with clusters in space and time, the proposed methodology is compared with the spatial version of Byers and Raftery (1998). The comparison is done in terms of the true positive rate, the false positive rate and the accuracy. In general, when weighting the temporal component in the distance measure, the results of the classification improve. In comparison to the spatial method, our methodology outperforms the other one. The analysis of the California sequences, to identify background and triggered events, shows its utility and wide application.

References

Adelfio G, Chiodi M (2015) Flp estimation of semi-parametric models for space–time point processes and diagnostic tools. Spatial Statistics 14:119 – 132

Allard D, Fraley C (1997) Nonparametric maximum likelihood estimation of features in spatial point processes using voronoi tessellation. Journal of the American Statistical Association 92(440):1485–1493

Byers S, Raftery AE (1998) Nearest-neighbor clutter removal for estimating features in spatial point processes. Journal of the American Statistical Association 93(442):577–584

Dasgupta A, Raftery AE (1998) Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American Statistical Association 93(441):294–302

Demattei C, Cucala L (2010) Multiple spatio-temporal cluster detection for case event data: An ordering-based approach. Communications in Statistics - Theory and Methods 40(2):358–372

Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical Analysis and Modelling of Spatial Point Patterns, vol 70. John Wiley & Sons

Mateu J, Lorenzo G, Porcu E (2007) Detecting features in spatial point processes with clutter via local indicators of spatial association. Journal of Computational and Graphical Statistics 16(4):968–990

Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association 83(401):9–27

Siino M, Rodríguez-Cortés FJ, Mateu J, Adelfio G (2019) Spatio-temporal classification in point patterns under the presence of clutter. Submitted

Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. Remote sensing of Environment 62(1):77–89