PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

GRASPA 2023

GRASPA-SIS BIENNAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

Palermo, 10-11 July, 2023

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

GRASPA

SIS Società Italiana di Statistica

TIES

dSEAS dipartimento scienze economiche aziendali e statistiche

Università degli Studi di Palermo

Proceedings of the GRASPA 2023 Conference
Palermo, 10-11 July 2023
Edited by: Giada Adelfio and Antonino Abbruzzo
–
Palermo: Università degli Studi di Palermo.

ISBN: 979-12-210-3389-2

Sponsored by:

Selecting the Kth nearest-neighbour for clutter removal in spatial point processes through segmented regression models

Nicoletta D'Angelo[1*] and Giada Adelfio[1]

[1] *Department of Economics, Business and Statistics, University of Palermo, Italy; nicoletta.dangelo@unipa.it, giada.adelfio@unipa.it*
[*] *Corresponding author: Nicoletta D'Angelo*

Abstract. *We consider the problem of feature detection, in the presence of clutter in spatial point processes. A previous study addresses the issue of the selection of the best nearest neighbour for clutter removal. We outline a simple workflow to automatically estimate the number of nearest neighbours by means of segmented regression models applied to an entropy measure of cluster separation. The method is suitable for a feature with clutter as two superimposed Poisson processes on any twodimensional space, including linear networks. We present simulations to illustrate the method and an application to the problem of seismic fault detection.*

Keywords. *Changepoint detection; Clutter; EM-Algorithm; Feature; Spatial point processes*

---

1 Introduction

One of the main interests of spatial point pattern analysis is identifying features surrounded by clutter. Byers and Raftery (1998) assume that the clutter is distributed as a homogeneous Poisson point process with some complex geometry. The features are also homogeneous Poisson point processes but with different intensities, restricted to a certain sub-region and overlaid on the clutter. Therefore, the resulting point process is Poisson with piece-wise constant intensity on the analysed region, and there are no assumptions about the shape of the features or their densities. In the spatio-temporal context, Siino et al. (2020) studies scenarios with multiple features and several shapes of the clusters. In this paper, we will consider both homogeneous Poisson processes and clustered Poisson processes.
The aim of this work is to provide different methods to deal with more complex scenarios. In detail:

- Automatically select the number of nearest neighbors to consider, by means of segmented regression models applied to an entropy measure of cluster separation;
- Application of the procedure multiple times to get a "better" end result. This would treat the estimated feature as a new dataset and reapply the method on it.

The structure of the paper is as follows. Section 2 presents the method for feature detection by Byers and Raftery (1998). Section 3 gives some basics about segmented regression models and introduces the proposed method. Section 4 shows a simulation study to assess its performance. Section 5 contains an application to the problem of detecting seismic faults. Section 6 concludes the paper.

2 Kth nearest neighbor distribution of distances

Let D_K be the distance of the Kth nearest neighbour of u; if D_K is greater than r_u, then there must be one of $0,1,\ldots,K-1$ points at a distance less than r_u. For all $u \in W$ and $x \in [0,\infty)$, the Kth nearest neighbour distribution approximation is given by

$$\mathbb{P}(D_K \geq x) = \sum_{k=0}^{K-1} \frac{e^{-\lambda\pi x^2}(\lambda\pi x^2)^k}{k!} = 1 - F_{D_K}(x),$$

where $\mathbb{P}(D_K \geq x)$ is the probability that the Kth nearest neighbour point falls out of $b(u,x)$ with $|b(u,x)| = x$, assuming that this disk exists around u. If the Kth nearest neighbour point of u is outside $b(u,x)$, it is also outside $b(u,r_u)$.
Accordingly, the density $f_{D_K}(x)$ can be found as

$$f_{D_K}(x) = \frac{e^{-\lambda\pi x^2} 2(\lambda\pi)^K x^{2K-1}}{(K-1)!}, \tag{1}$$

and therefore $Y \sim \Gamma(K,\lambda\pi)$, with $Y = (D_K)^2$. Having a closed-form and the Gamma distribution properties, the maximum likelihood estimation of the rate given the observed values of D_K is straightforward. Indeed, the maximum likelihood estimate of λ is

$$\hat{\lambda} = \frac{nK}{\pi \sum_{i=1}^{n} d_i^2},$$

where d_i is the ith observed Kth nearest neighbour distance.
We assume two types of processes to be classified through a mixture of the corresponding Kth nearest neighbour distances coming from the clutter and feature, which are two superimposed Poisson processes. Therefore, based on equation (1), we assume that

$$D_K \sim p\Gamma^{1/2}(K,\lambda_1\pi) + (1-p)\Gamma^{1/2}(K,\lambda_2\pi), \tag{2}$$

where λ_1 and λ_2 are the intensities of the two homogeneous Poisson point processes (clutter and feature) and p is the constant that characterizes the postulated distribution of the D_K.
The parameters λ_1, λ_2, and p associated with the mixture are estimated using an EM algorithm, wherein we use the closed-form of a Gamma distribution in the expectation step. On the other hand, let $\delta_i \in \{0,1\}$ be the two classification components for each data point, where $\delta_i = 1$ if the ith point belongs to the feature and $\delta_i = 0$ otherwise. Thus each data point has an observation d_i of D_K and an unknown δ_i. Hence, the $\mathbb{E}$ step of the algorithm consists of

$$\mathbb{E}[\hat{\delta}_i^{(t+1)}] = \frac{\hat{p}^{(t)} f_{D_K}(d_i;\hat{\lambda}_1^{(t)})}{\hat{p}^{(t)} f_{D_K}(d_i;\hat{\lambda}_1^{(t)}) + (1-\hat{p}^{(t)}) f_{D_K}(d_i;\hat{\lambda}_2^{(t)})},$$

and the maximization M step consists of

$$\hat{\lambda}_1^{(t+1)} = \frac{K\sum_{i=1}^{n} \hat{\delta}_i^{(t+1)}}{\pi \sum_{i=1}^{n} d_i^2 \hat{\delta}_i^{(t+1)}}, \quad \hat{\lambda}_2^{(t+1)} = \frac{K\sum_{i=1}^{n} (1-\hat{\delta}_i^{(t+1)})}{\pi \sum_{i=1}^{n} d_i^2 (1-\hat{\delta}_i^{(t+1)})}, \quad \hat{p}^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{\delta}_i^{(t+1)}}{n}.$$

An intuitive classification test criterion would classify the points according to the mixture component where the distances have the highest densities. We are mainly interested in identifying the feature points in this proposed classification approach; consequently, we do not consider the edge effects because feature points in practice are predominantly far from the edges. Additionally, for large n, the convergence of the EM algorithm is good, since it takes less time to arrive at an approximately acceptable solution, and it does so with the fewest number of iterations.

3 Selecting K through changepoint detection

Segmented or broken-line models are regression models where the relationships between the response and one or more explanatory variables are piecewise linear and, as such, represented by two or more straight lines connected at unknown points. These models are a common tool in many fields, including epidemiology, occupational medicine, toxicology, and ecology, where it is usually of interest to assess threshold values where the effect of the covariate changes. The main advantage of this approach is the easy interpretation given by two components, i.e. changepoints and slopes. The segmented linear regression is expressed as

$$g(E[Y|x_i,z_i]) = \alpha + z_i^T\theta + \beta x_i + \sum_{k=1}^{K_0}\delta_k(x_{i,k}-\psi_k)_+ \tag{3}$$

where g is the link function, x_i is the broken-line covariate and z_i is a covariate vector whose relationship with the response variable is a non-broken line. We denote by K_0 the true number of changepoints and by ψ_k the K_0 locations of the changepoints in the observed phenomenon. These K_0 are selected among all the possible values in the range of x. The term $(x_i-\psi_k)_+$ is defined as $\sum_i I(x_i>\psi_k)$ that is $(x_i-\psi_k)I(x_i>\psi_k)$. The parameter estimates $\boldsymbol{\theta}$ represent the non broken-line effects of z_i, β represents the effect for $x_i<\psi_1$, while $\boldsymbol{\delta}$ is the vector of the differences in the effects. The parameters to be estimated are the number of changepoints K_0; their locations ψ_k; and the broken-line effects, represented by β and $\boldsymbol{\delta}$ (Muggeo, 2003).

The development mentioned in Section 2 assumes a proper value of K priorly chosen. The natural way to choose the suitable Kth neighbour is by analysing several increasing values of K and then selecting the K after which no improvement is found. However, in the literature, there are several methodological proposals for this target; in this work, we use an entropy-type measure of separation introduced in Celeux and Soromenho (1996) given by

$$S = -\sum_{i=1}^{n}\delta_i\log_2(\delta_i), \tag{4}$$

where δ_i are the probabilities of being in the first component of the mixture in equation (2) which is the feature. As stated by Byers and Raftery (1998), plotting the entropies sequentially and looking for a levelling-off changepoint in the graph is an easy way to choose K. Therefore, we propose to select the optimal K by fitting a segmented model as in Diaz-Sepulveda et al. (2022), by fitting

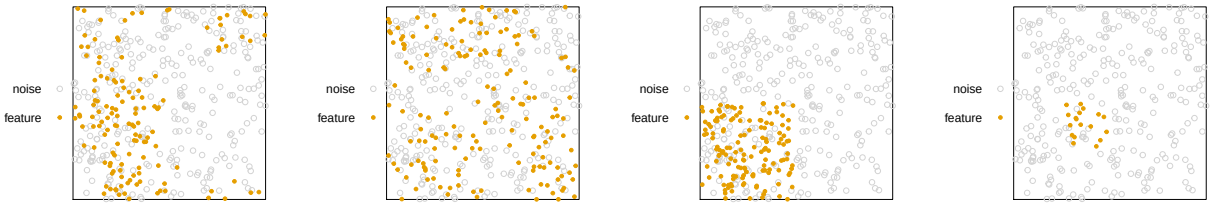
$$\mathbb{E}[Y|x_i] = \beta + \delta(x_i-\psi)I(x_i>\psi), \tag{5}$$

where the interest is estimating a unique changepoint ψ, after which the slope $\beta+\delta$ is constrained to be equal to zero. The observed response variable Y is the entropy level in (4), modelled as a function of the number of nearest neighbours, the covariate x. The following steps implement the classification procedure:

(1) Choose a value of K either by imputing a sought value or automatically through the application of a segmented regression model;
(2) Find the Kth nearest neighbours distance for each point in the point pattern;
(3) Apply the EM algorithm for estimating λ_1, λ_2, and p;
(4) Classify the points according to whether they have a higher density under the feature or clutter component of the mixture;
(5) Repeat steps (1) - (4) iteratively until some stopping criterion is met.

4 Simulations

This section aims to study the proposed method's performance in terms of classification rates considering different scenarios, concerning both the generating processes and the ratio between the number of clutter and feature points generated. To this end, we simulate different such scenarios, to obtain a comprehensive understanding of the results of the proposed algorithm in different settings.
Examples of the simulated patterns are depicted in Figure 1. In detail, we simulate 200 clutter homogeneous Poisson point processes with expected number $\mathbb{E}[n_c]$. The feature point patterns are simulated in the sub-window W_f from the following processes: (a-b) homogeneous Poisson cluster process, which are commonly used to model spatially clustered patterns, with a general structure considering Poisson point process of cluster centers and a random number of points distributed about each cluster center. We denote with κ the intensity of the Poisson process of cluster centers and v the number of points constituting each cluster in a disc of radius 0.2; (c-d) homogeneous Poisson processes.



(a) (b) (c) (d)

Figure 1: Examples of simulated patterns of Table 1.

We show the results of the proposed procedure in Table 1 in terms of true-positive rate (TPR), false-positive rate (FPR), and accuracy (ACC), averaging over the simulated point patterns. We run the proposed algorithm for fixed numbers of $K = \{10, 20, 30\}$, and with the one estimated by equation 5. Moreover, we consider up to three iterations, to assess whether this helps to get better end results.
The rates are defined as

$$TPR = \frac{\text{true positives}}{\text{positives}}, \quad FPR = \frac{\text{false negatives}}{\text{negatives}}, \quad ACC = \frac{\text{true positives and negatives}}{\text{positives and negatives}}.$$

Obviously, it is desirable having both TPR and ACC close to 1 and FPR close to 0.
Increasing the fixed K, the TPR increases, but the FPR increases as well. Also, the accuracy decreases. Increasing the number of iterations with the estimated $\hat{K}$, the TPR decreases, as of course we are restricting the possible feature points at each iteration. Selecting $\hat{K}$ with segmented regression leads to better results in terms of accuracy, as the number of iterations increases.

Table 1: Classification rates averaged over 200 point patterns simulated on the unit square with $\mathbb{E}[n_c]$ and $\mathbb{E}[n_f]$ expected number of points for clutter and feature.

					K		iter 1	K		iter 2	$\hat{K}$	
	Feature process	$\mathbb{E}[n_c]$	$\mathbb{E}[n_f]$	Rate	10	20	30	10	20	30	iter 1	iter 2
(a)	Poisson cluster	300	150	TPR	0.778	0.790	0.783	0.601	0.605	0.596	0.789	0.572
	$W_f = [0,1]$			FPR	0.584	0.592	0.604	0.370	0.375	0.385	0.592	0.342
	$\kappa = 7.5 \quad v = 20$			ACC	0.569	0.547	0.524	0.611	0.612	0.602	0.549	0.602
(b)	Poisson cluster	300	150	TPR	0.737	0.746	0.739	0.588	0.544	0.543	0.745	0.531
	$W_f = [0,1]$			FPR	0.630	0.640	0.647	0.450	0.411	0.424	0.639	0.395
	$\kappa = 15 \quad v = 10$			ACC	0.524	0.487	0.468	0.574	0.558	0.553	0.490	0.554
(c)	Poisson	300	150	TPR	0.960	0.970	0.972	0.708	0.727	0.732	0.968	0.699
	$W_f = [0,0.5]$			FPR	0.429	0.419	0.427	0.068	0.087	0.111	0.437	0.060
				ACC	0.742	0.753	0.746	0.783	0.788	0.783	0.736	0.779
(d)	Poisson	300	20	TPR	0.996	1.000	1.000	0.635	0.617	0.637	0.999	0.591
	$W_f = [0.25,0.5]$			FPR	0.885	0.897	0.903	0.056	0.010	0.001	0.894	0.014
				ACC	0.518	0.440	0.418	0.654	0.639	0.658	0.452	0.614

5 Detecting seismic faults

Dasgupta and Raftery (1998) considered the problem of detecting seismic faults based on an earthquake catalog, to show the performance of their proposed MClust-EM algorithm. The idea is that earthquake epicenters occur along seismically active faults and are measured with some error. So over time, observed earthquake epicenters should be clustered along such faults. The analysed earthquake catalog was recorded over a 40,000 km^2 region of the central coast ranges in California from 1962-1981 (McKenzie et al., 1982). This region is characterized by a well-documented known fault structure, so it is easier to compare the obtained and the expected results.

Dasgupta and Raftery (1998) selected a classification with seven clusters (six nonnoise clusters and one noise cluster), because the BIC attains a local maximum there and the successive differences in the BIC values are small thereafter. They found that the classification obtained using six (nonnoise) clusters corresponds well with the available documentation of faults in the region of interest. One or two clusters do not correspond to any of the documented faults.

An application of 5th NN clutter removal produced the results on Byers and Raftery (1998). One key difference is the isolated cluster in the bottom right that NN methods pick up but that the connected component part of Allard and Fraley's method leaves out. This cluster is treated as one end of a linear cluster of earthquakes in the analysis of Dasgupta and Raftery (1998). They end up filling in the sparse part between it and other clusters with clutter to produce the linear form that they search for. It would seem that the MClust-EM method is more suited to finding features, such as faults that are supposed to be roughly linear. Otherwise, less-structured methods, both in terms of number and shape of features, like the Byers and Raftery (1998)'s, still achieve good results when dealing with "structured" situations.

We analyse the same catalog of the Northern California earthquakes, with magnitude at least 2.5, available from `https://ncedc.org/ncedc/catalog-search.html`. As the dataset contains duplicated points, we select 1399 non-overlapping points. We proceed to run the proposed iterative procedure up to three iterations. The nearest neighbors selected at each iteration are 3, 5, and 4. The detected feature points are displayed in Figure 2.

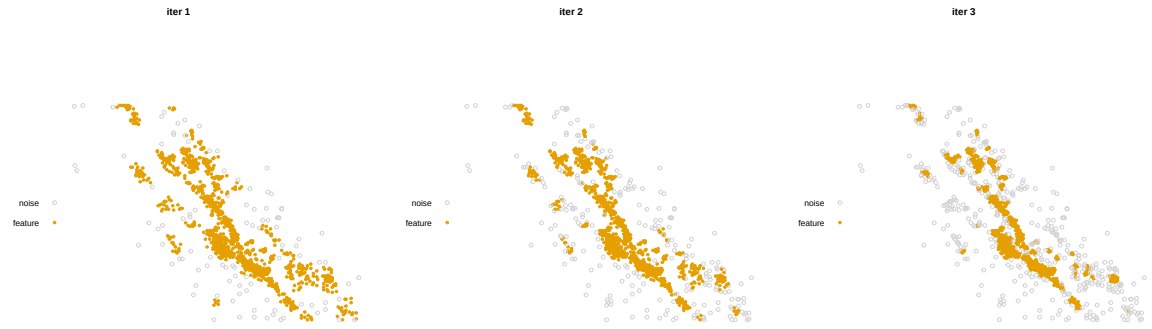


Figure 2: Output of the proposed iterative procedure up to 3 iterations, applied to the analysed earthquake data.

6 Conclusions

We have addressed the problem of the selection of the Kth nearest-neighbour in clutter removal problems for spatial point processes. A criterion to select the best K is crucial when analysing real data. At this aim, we have proposed an automatic selection procedure by means of the segmented regression models and assessed its performance through simulations. Results have shown that our automatic proposal, as well as iteratively applying the procedure to the previously labeled feature points, performs similarly to the benchmark methodology, in terms of accuracy. Therefore, we suggest using our proposed selection method in applications, when of course the best K is not known in advance.
Future work includes the definition of a stopping criterion for the iterative procedure and an extended simulation study and some applications in both the Euclidean and the linear network context.

Acknowledgments. This work was funded by 'FFR 2023 GIADA ADELFIO', 'FFR 2023 NICOLETTA D'ANGELO', and by the PNRR project, grant agreement No PE0000018 - GRINS.

References

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

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13(2):195–212.

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

Diaz-Sepulveda, J. F., D'Angelo, N., Adelfio, G., Gonzalez, J. and Rodriguez-Cortez, F. J. (2022). Nearest-Neighbor Clutter Removal for Estimating Features in Linear Point Processes. *Preprint*, https://arxiv.org/abs/2209.14082.

McKenzie, M., Miller, R., and Uhrhammer, R. (1982). *Bulletin of the seismographic stations*. University of California, Berkeley, 53(1-2).

Muggeo, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071.

Siino, M., Rodríguez-Cortés, F.J., Mateu, J. and Adelfio, G. (2020). Spatio-temporal classification in point patterns under the presence of clutter *Environmetrics*, 31(2), e2599.