Clustering in point processes on linear networks using nearest neighbour volumes

Juan F. Díaz-Sepúlveda jfdiazs0@unal.edu.co* [1], Nicoletta D'Angelo[2], Giada Adelfio[2], Jonatan A. González[3] and Francisco J. Rodríguez-Cortés[1]

[1]Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia
[2]Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Palermo, Italy
[3]Department of Mathematics, University Jaume I, Castellon, Spain

Abstract

In this study, we address the challenge of detecting clusters within linear networks, focusing on the classification of point processes. We extend the classification method developed in previous studies to this more complex geometric context, where the classical properties of a point process change and data visualization are not intuitive. Our approach leverages the distribution of the K-th nearest neighbour volumes in linear networks. Consequently, our methodology is well-suited for analysing point patterns comprising two overlapping Poisson processes occurring on the same linear network. To illustrate the method, we present simulations and examples of road traffic accidents that resulted in injuries or deaths in two cities in Colombia.

Keywords– Clutter, EM Algorithm, Feature, Kth nearest-neighbour, Linear network, Spatial point pattern.

1 Introduction

In the last decade, spatial statistics has undergone an extraordinary methodological and computational advancement focused on generalising and extending the foundations of the theories to more complex geometric spaces that allow fairer statistical analysis of new kinds of spatial data. In particular, spatial point process methodologies have been addressed to consider non-classical geometric supports for analysing events on linear networks, such as traffic accidents in a city (Baddeley et al., 2021a). One of the most relevant contributions to

*Corresponding author: jfdiazs0@unal.edu.co (Juan F. Díaz-Sepúlveda).

the statistical analysis of network data is the definition of the geometrically corrected first- and second-order characteristics. These descriptors can be used to identify spatial configurations and discriminate between point patterns according to their aggregation structure (Ang et al., 2012; McSwiggan et al., 2017; Moradi et al., 2018; Rakshit et al., 2019a,b; D'Angelo et al., 2021, 2023), select or reduce a set of variables (Rakshit et al., 2021) and estimate relative risks (McSwiggan et al., 2020).

Along with this work, we want to provide a statistical mechanism that can uncover hidden clusters within linear network data. We use the definition of cluster used in density-based clustering, whose basic idea is that the data which is in the region with high density of the data space is considered to belong to the same cluster (Kriegel et al., 2011; Xu and Tian, 2015). Kriegel et al. (2011) say that density-based clusters can be referred to as "natural clusters" since they are suitable for certain nature-inspired applications, as for example, in spatial data, clusters of points can be formed along natural structures such as rivers, roads, seismic faults, etc. In epidemiology, for instance, identifying disease clusters along transportation routes can help pinpoint the origins of outbreaks and implement timely interventions. In transportation planning, detecting clusters of traffic accidents or congestion on road networks is crucial for improving safety and optimizing infrastructure. Moreover, understanding how specific events cluster in linear networks can inform decision-making and policy formulation in urban design and environmental science. To this end, we depart from exploring classifying events (data points) coming from different point processes. For example, detecting surface minefields from images of reconnaissance aircraft can be processed to obtain a list of objects, some of which may be mines and other objects (Allard and Fraley, 1997; Byers and Raftery, 1998). For spatial point processes, this problem has been addressed differently. Allard and Fraley (1997) developed a method to find the maximum likelihood solution using Voronoi polygons. Dasgupta and Raftery (1998) used model-based clustering to extend the methodology proposed by Banfield and Raftery (1993). While these methods are based on some limiting assumptions, Byers and Raftery (1998) adopted a different approach by classifying the mines and the other objects with no assumptions about the shape or number of mines. More recently, González et al. (2021) consider the local contributions of the pair correlation function as functional data and propose two classification procedures for the two point patterns.

When minefields are distributed over the streets of a city or a neighbourhood, i.e., over a linear network, the classification problem cannot be analysed similarly because the events do not occur in the whole plane. Indeed, Baddeley et al. (2021a) states that statistical analysis of point data on linear networks presents severe problems because the associated graph is not spatially homogeneous, creating geometric and computational complexities and leading to new problems with a high risk of methodological error. In addition, data on networks can have a wide range of spatial scales. Baddeley et al. (2021a) emphasise that the point processes on linear networks also challenge the classical methodology of spatial statistics based on stationary processes, which is mainly inapplicable to data on linear networks.

This paper presents a novel extension of the spatial technique originally introduced by

Byers and Raftery (1998), which allows estimating regions of differing point densities in a point process. In our work, we advance this technique from the planar domain to the context of linear networks. This extension aims to provide a robust classification method tailored to detect clustering, where clustering is defined as regions within the linear network where the expected point density of a point process is elevated. The application of this method is particularly interesting in analysing traffic accidents, where identifying and understanding clusters along linear networks can greatly enhance our ability to enhance road safety, optimize traffic management, and make informed decisions for accident prevention and mitigation.

As the distribution of K-th nearest neighbour distances proposed by Byers and Raftery (1998) is not directly applicable to linear networks, we use the K-th nearest neighbour volumes defined by the K-th nearest neighbour distances. Following Byers and Raftery (1998), these volumes are modelled as a mixture distribution. The parameters are estimated by an EM algorithm to classify data points as two different types of point processes on the linear network.

The structure of the paper is as follows. Section 2 gives some basic techniques for analysing point patterns on linear networks. Section 3 presents the proposed method to classify cluster as overlapping Poisson processes on linear networks. Section 4 shows an extended simulation study carried out with different linear networks. Section 5 contains two applications of the classification method for traffic accidents in Bogota and Medellin (Colombia). Section 6 presents some conclusions.

2 Mathematical framework

Following (Ang et al., 2012; Baddeley et al., 2021a), we consider a linear network as the union of a finite number of line segments on the plane, namely $L = \bigcup_{h=1}^{m} l_h$, such that each $l_h = [u_h, v_h] = \{w : w = tu_h + (1-t)v_h, 0 \leq t \leq 1\}$, and u_h, v_h are the endpoints of the segment l_h for $h = 1, \ldots, m$. The total length of the segments contained in the network L, defined as $|L|$, is denoted by the term volume (Ang et al., 2012; Baddeley et al., 2021a), hereafter. Additionally, a path between two points u and v in a linear network L is a sequence $y_0, y_1, \ldots, y_q$ of points in L so that $y_0 = u$, $y_q = v$ and $[y_j, y_{j+1}] \subset L$ for each $j = 0, \ldots, q-1$. The path length is the sum of the lengths of its segments. The shortest path distance $d_L(u, v)$ between u and v in L is the minimum of the lengths of all paths from u to v. If there are no paths from u to v (implying that the network is not connected), then the distance is defined by $d_L(u, v) = \infty$.

We also assume that the disc of radius $r > 0$ and centre point $u \in L$ is the set of all points v in the network lying no more than a distance r from u, that is $b_L(u, r) = \{v \in L : d_L(u, v) \leq r\}$. The relative boundary of the disc is the set of points lying exactly r units away from u, $\partial b_L(u, r) = \{v \in L : d_L(u, v) = r\}$. The circumference $m(u, r)$ is the number of points of L that fall in $\partial b_L(u, r)$. The circumference is finite for all $r < \infty$, and we set $m(u, \infty) = \infty$ by convention. The circumradius of the network is the radius of the smallest

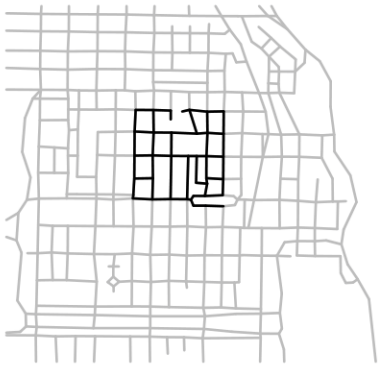
disc that contains the entire network, $R = R(L) = \inf\{r : m(u,r) = 0, \text{ for some } u \in L\} = \min_{u\in L} \max_{v\in L} d_L(u,v)$. If L is not path-connected, then R is the minimum of the circumradius of the connected components of L; for more details see (Ang et al., 2012; Baddeley et al., 2021a).

We define a (finite, simple) point pattern X on L as a finite set $X = \{x_1, x_2, \ldots, x_n\}$ of distinct points $x_i \in L$, which is the concrete realisation of a point process, where n is a positive integer (see, e.g., Rakshit et al., 2021). Further, $N_X(B) = N(X \cap B)$ is the number of points on X lying in B for any set $B \subset L$. Thus, for a finite and simple point possess with intensity function $\lambda(u)$, $u \in L$, we have $\mathbb{E}[N_X(B)] = \Lambda(B) = \int_B \lambda(u)\mathrm{d}_1 u$, for all measurable $B \subset L$, where $\mathrm{d}_1 u$ denotes integration with respect to arc length (Federer, 1969; Ang et al., 2012). If λ is constant, then X is called homogeneous, and the constant λ is the mean number of points per unit length for a homogeneous point process. A homogeneous Poisson point process on L with intensity $\lambda > 0$ is characterized by the properties that, for any line segment $B \subset L$, the number of points falling in B has a Poisson distribution with mean $\lambda|B|$. In contrast, events occurring in disjoint line segments $B_1, \ldots, B_k \subset L$ are independent for k a positive integer (Ang et al., 2012). A sub-network of L is a linear network, a subset of L. Finally, we define a cluster of points on a linear network as regions of high-point density that are separated by regions of low point density and, thus, can be arbitrarily shaped in the data on the network (Kriegel et al., 2011).

3 Clustering detection approach for point processes on linear networks

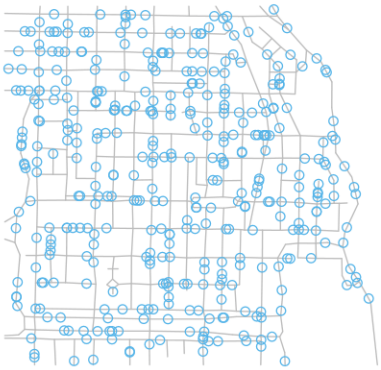
In spatial point pattern analysis, some authors have been interested in identifying clusters of points. Unlike the planar case, which is very rarely straightforward, identifying clusters in point processes on a linear network through visual inspection is more difficult due to the complexity of the domain. We consider the extension of the approach in Byers and Raftery (1998) by generating two point patterns on a linear network such that the first point pattern is a realisation of a homogeneous Poisson point process with intensity λ_1, restricted to a certain sub-network. The second point pattern is also a realisation of a homogeneous Poisson point process on the entire linear network with some intensity λ_2 and overlaid on the first point pattern. Therefore, the resulting point process is Poisson with piece-wise constant intensity on the linear network, which results in a higher density of points in the sub-network obtaining a cluster of points.

In order to illustrate the geometric context for our developments, we consider the linear network `chicago` of the package `spatstat.data` (Baddeley et al., 2021b), which represents the street network of some area of Chicago (Illinois, USA) close to the University of Chicago; a particular case of simulated data on this linear network is shown in Figure 1 at four different stages (Ang et al., 2012). The sub-network where the first point pattern is generated is shown in Figure 1a. The street network has 338 intersections, 503 uninterrupted segments and a

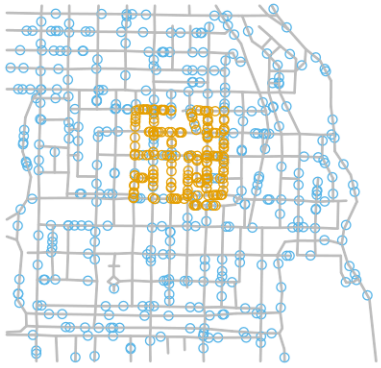


(a) Sub-network of `chicago` for the first point pattern.

(b) First point pattern.



(c) Second point pattern.



(d) Superposition of the first and second point patterns.

Figure 1: Linear network `chicago` at four different stages. A first point pattern simulated from a homogeneous Poisson point process with $\lambda_1 = 0.067$ and $\mathbb{E}[n_1] = 200$. The second point pattern is an observation of a homogeneous Poisson point process with $\lambda_2 = 0.013$ and $\mathbb{E}[n_2] = 400$.

total length of 31150 feet. The sub-network has 39 intersections, 53 uninterrupted segments and a total length of 2991 feet. Figure 1b shows the first point pattern simulated from a Poisson process with intensity $\lambda_1 = 0.067$ and expected number of points $\mathbb{E}[n_1] = 200$ in the sub-network. Figure 1c shows the second point pattern simulated from a homogeneous Poisson point process with intensity $\lambda_2 = 0.013$ and expected number of points $\mathbb{E}[n_2] = 400$ across the entire linear network. The superposition of the first and second point patterns in Figures 1b and 1c are displayed in Figure 1d, which is the resulting point pattern, that is, what we typically observe.

3.1 K-th nearest neighbour volumes distribution

If the number of independent random events occurring in a region of a specific area is a Poisson variable with constant intensity λ, then for an integer α, the required area for the occurrence of α-th Poisson events in a sub-region has a Gamma distribution, $\Gamma(\alpha, \lambda)$, where shape and rate parameters are α and λ, respectively (Mood et al., 1974). In the same way, it can be said that if the number of independent random events occurring in a linear network is a Poisson variable with constant intensity λ, the sub-network volume that is necessary for α-th Poisson events to fall within this sub-network has a distribution $\Gamma(\alpha, \lambda)$.

In the linear networks geometry, the disc $b_L(u, r)$ has variable volume for a given r and different locations u (i.e. the disc volume depends on u), as mentioned in Cronie et al. (2020). Assuming a Poisson process X on L with constant intensity λ, the number of points $N_X(b_L(u, r))$ of X falling in the disc $b_L(u, r)$, has an approximate Poisson distribution with mean equal to $\lambda|b_L(u, r)|$. Because $N_X(b_L(u, r))$ is truncated above by $|L|$, the Poisson distribution is an approximation. For $u \in L$, let $D_K^L(u)$ be the distance of the K-th nearest neighbour of u and $S_K(u) = |b_L(u, D_K^L(u))|$. As $N_X(b_L(u, D_K^L(u))) = K$, then $S_K = S_K(u)$ has an approximate Gamma distribution $\Gamma(K, \lambda)$.

Accordingly, the density of K-th nearest neighbour volume S_K is

$$f_{S_K}(s) = \frac{\lambda^K \; s^{K-1} \; e^{-\lambda s}}{\Gamma(K)}, \quad \text{for } s > 0; \; K, \lambda > 0. \tag{1}$$

Having a closed-form and the Gamma distribution properties, the maximum likelihood estimation of the rate given the observed values of S_K is straightforward. Indeed, the maximum likelihood estimate of λ is

$$\hat{\lambda} = \frac{nK}{\sum_{i=1}^{n} s_i} = \frac{K}{\bar{s}_i},$$

where $s_i = |b_L(x_i, D_K^L(x_i))|$ with $x_1, \ldots, x_n$ are points of the Poisson process X.

3.2 Mixture modeling and estimation

Six estimated volume distributions S_K for the mixture of a couple of homogeneous Poisson point patterns simulated in the `chicago` network (see Figure 1) from the 27-th to the 32-nd nearest neighbour are displayed as histograms in Figure 2, where a clear bimodality is evident. They can be used as exploratory tools and motivation to propose a model (Byers and Raftery, 1998). We assume two types of processes to be classified through a mixture of the corresponding K-th nearest neighbour volumes coming from the two superimposed Poisson processes as defined above. Based on equation (1), we assume that

$$S_K \sim p\Gamma(K, \lambda_1) + (1 - p)\Gamma(K, \lambda_2), \tag{2}$$

where λ_1 is the intensity of the first point pattern simulated in the sub-network, λ_2 is the intensity of the second point pattern simulated on the entire linear network, and p is the

constant that characterizes the postulated distribution of the volumes S_K. We intuitively assume that the first point pattern is part of the first component.

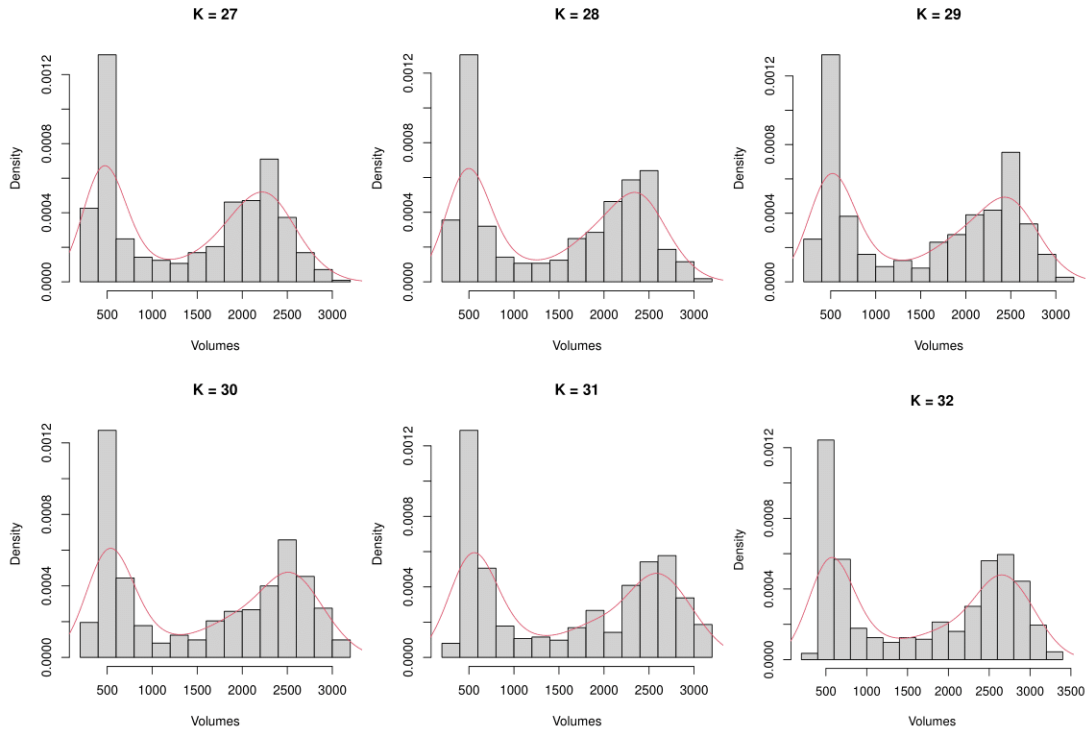


Figure 2: Distributions of the volumes S_K for the mixture of the overlapping point patterns from 27-th to 32-nd nearest neighbours simulated on the `chicago` network in Figure 1.

The parameters λ_1, λ_2 and p associated with the mixture are estimated using an EM algorithm (Dempster et al., 1977), 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 i-th point belongs to the first point pattern and $\delta_i = 0$ otherwise. Thus, each data point has an observed volume s_i of S_K and an unknown classification component δ_i. Hence, the E step of the algorithm consists of

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

and the maximization M step consists of

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

We follow an intuitive classification test criterion, classifying the points according to the mixture component where the volumes have the highest densities, that is, where the cluster

is located. We are mainly interested in identifying the two point patterns, and this proposed classification approach takes the estimated component with the highest intensity as the distribution of the first point pattern. Additionally, for large n, the convergence of the EM algorithm is good since it arrives at an approximately acceptable solution.

The aforementioned development relies on the assumption of a pre-selected appropriate value for K. A common approach to choosing the K-th neighbour involves evaluating various increasing values of K and selecting the point where no further improvement is observed. Nevertheless, existing literature offers several methodological suggestions for this purpose. In this study, we follow what proposed in D'Angelo and Adelfio (2023) for the Euclidean case, adopting the entropy-type measure of separation

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

where δ_i represents the probabilities of belonging to the initial component of the mixture in equation (2), corresponding to the first point pattern generated in the sub-network.

According to Byers and Raftery (1998), a practical strategy for selecting K involves plotting entropies sequentially and identifying a changepoint where the graph levels off. Figure 3 illustrates this process, displaying classification entropies for K values up to 35, obtained from the simulated point pattern on the `chicago` network shown in Figure 1.

To automate the selection of the optimal K, we utilize a segmented regression model

$$\mathbb{E}\left[Y|x_i\right] = \beta + \gamma(x_i - \psi)I(x_i < \psi),$$

where we aim to estimate a single changepoint ψ. After this point, the slope is constrained to be zero. As illustrated in Figure 3, the observed response variable Y represents the entropy level and is modelled as a function of the number of nearest neighbours x. For further details on inferential aspects of segmented regression models, we refer readers to Muggeo (2003).

The following steps implement the classification procedure:

(1) Choose a value of K either by imputing a sought value or automatically applying a segmented regression model. Note that an upper bound for the K domain should be fixed manually to a reasonable maximum number of neighbours.

(2) Find the K-th nearest neighbour distance and calculate the disc volume for each point in the point pattern using the shortest path distances computed on the given linear network.

(3) Apply the EM algorithm for estimating λ_1, λ_2, and p.

(4) Classify the points in the first point pattern according to whether they have a higher density under the mixture's components and take these as points belonging to the cluster.

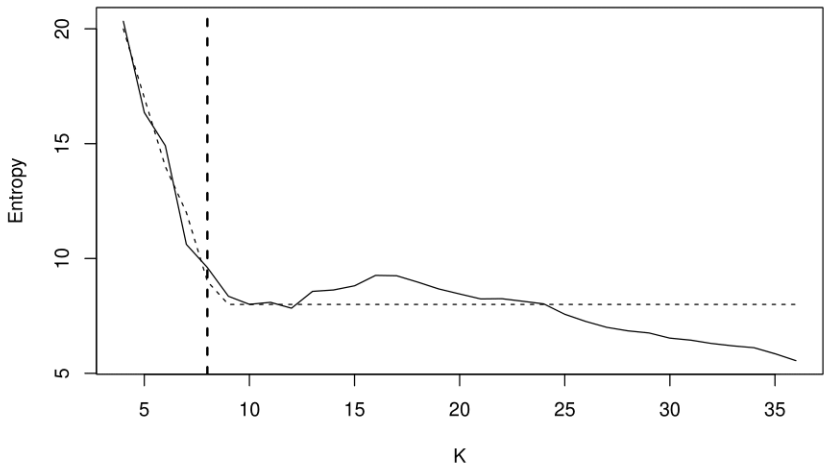


Figure 3: Estimated K values for the simulated pattern on the `chicago` network in Figure 1. The black line represents the observed entropies, and the dotted line represents the estimated segmented model. The vertical line indicates the estimated changepoint of $\hat{K} = 8$.

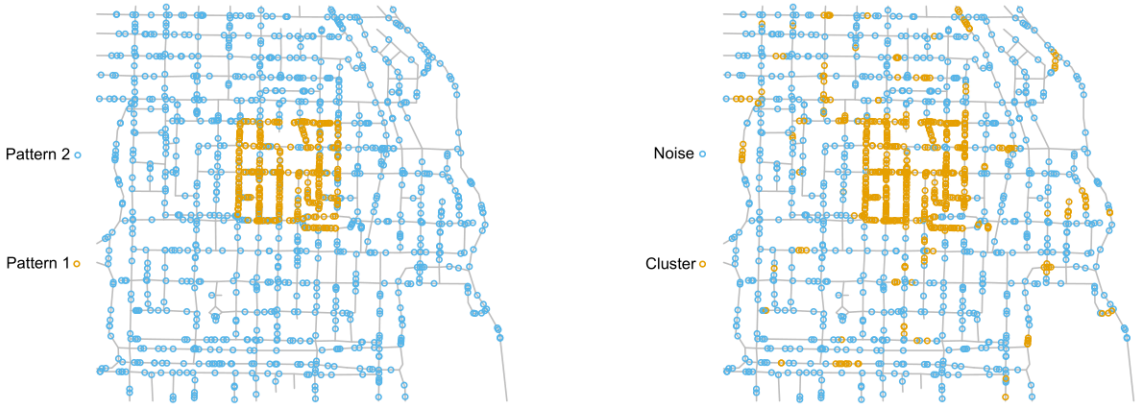
4 Simulation study

This section aims to study the proposed method's performance in terms of clustering detection rates, considering different scenarios for both point patterns and shapes of sub-networks. We simulate the first point pattern constrained to a sub-network and generate the second point pattern in the whole linear network in all scenarios. To explore the classification procedure for different networks, we use both the `chicago` and the `dendrite` networks (Baddeley et al., 2021b) in the simulation study, which represents linear networks with and without cycles. We use the `chicago` network in the third scenario, taking two nested sub-networks and generating the second point pattern with two different intensities. The fourth network is the Antonio Narino network (Moncada, 2018), which has a homogeneous structure, and we simulate the first point pattern in two disjoint sub-networks corresponding to two separate roads.

We show the results of the clustering method in terms of true-positive rate (TPR), false-positive rate (FPR), and accuracy (ACC), averaging over 100 simulated point patterns generated with $\mathbb{E}[n_1]$ and $\mathbb{E}[n_2]$ expected number of points for the first and second point patterns, respectively. In addition, the λ_1 and λ_2 intensities are reported for each pattern. The accuracy is the proportion of correct predictions (both true positives and negatives) among the total number of cases examined. The tables also report the average of the 100 estimates of K by the segmented regression model for the simulated point patterns in each design ($\bar{K}$) and its standard deviation (sd). We further compare the detection with K equal to $\{5, 10\}$ and the automatically selected by the segmented regression model. The rates are computed accordingly.

4.1 Clustering detection on a linear network with cycles

The linear network `chicago` described in Section 3 is a network with loops given with edges connecting vertexes to themselves. As mentioned in that section, we generate two point patterns as displayed in Figure 4a, and report the outcomes in Table 1.



(a) Simulate pattern. (b) Clustering pattern.

Figure 4: *(a):* Simulated point pattern on the `chicago` linear network. The first and second patterns are simulated from Poisson point processes homogeneous with $\lambda_1 = 0.100$ and $\mathbb{E}[n_1] = 300$ and $\lambda_2 = 0.032$, $\mathbb{E}[n_2] = 1000$, respectively. *(b):* Randomly selected classified point pattern using the proposed clustering method.

Designs 1 and 2, where $\lambda_1 > \lambda_2$, show good performance in true and false positive rates for $K = 5$, $K = 10$ and $\hat{K}$. For designs 3 and 4, where $\lambda_1 \approx \lambda_2$, the clustering method performs well in true-positive rates for $K = 5$, $K = 10$ and $\hat{K}$, but there is a significant increase in false-positive rates. In designs 5 and 6, where $\lambda_1 < \lambda_2$, the clustering method has lower performance for $K = 5$, $K = 10$ and $\hat{K}$, with a decrease in the true-positive rates and an increase in the false-positive rates. The True Positive rates always being higher than false-positive rates. The accuracy shows that the best performance is achieved when the expected number of points of the first point pattern approaches the expected number of points of the second point pattern when their ratio comes close to 1. The worst performance is achieved when this ratio is close to 0. In the former case, there are few points of the second pattern compared to the number of points of the first pattern, and the clustering method quickly detects the points of the first pattern as in design 1. In this case the cluster has more intensity. In contrast, in the latter case, the amount of points in the second point pattern is too large, and the points in the first pattern are more complicated to detect, as in design 6. In this case the cluster is more hidden. The estimated $\hat{K}$ selected by modelling the entropy level has similar classification rates as with $K = 5$ and $K = 10$ for every design showing that values of K within a reasonable range, tend to perform similarly.

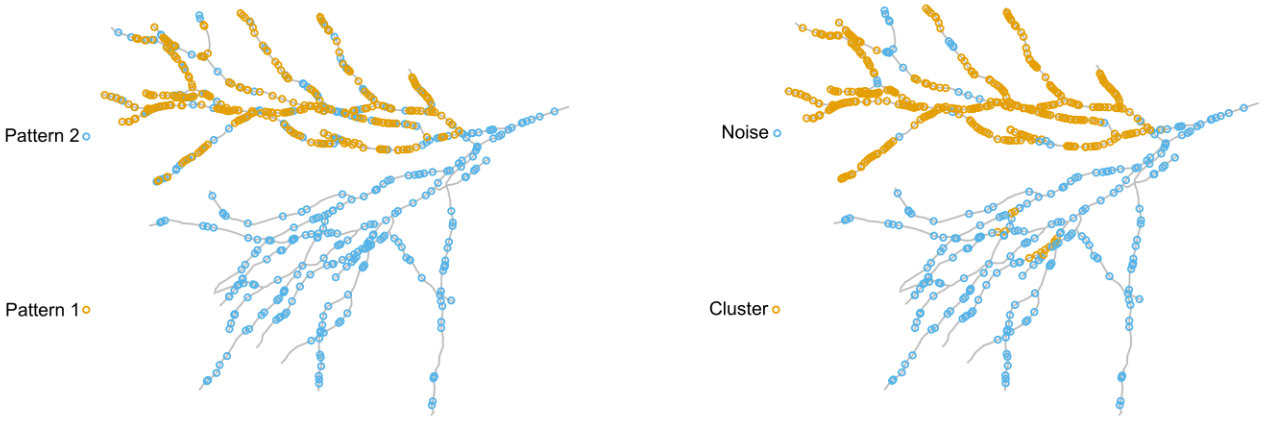
								K		
Design	λ_1	λ_2	$\mathbb{E}[n_1]$	$\mathbb{E}[n_2]$	$\bar{K}$	sd(K)	Rate	5	10	$\hat{K}$
1	0.100	0.032	300	1000	4.19	0.42	TPR	0.962	0.952	0.963
							FPR	0.241	0.134	0.274
							ACC	0.805	0.886	0.780
2	0.067	0.032	200	1000	4.81	0.49	TPR	0.944	0.931	0.946
							FPR	0.356	0.189	0.366
							ACC	0.694	0.831	0.686
3	0.033	0.032	100	1000	7.34	1.62	TPR	0.896	0.910	0.900
							FPR	0.490	0.398	0.436
							ACC	0.545	0.629	0.594
4	0.017	0.016	50	500	7.26	1.15	TPR	0.904	0.891	0.905
							FPR	0.502	0.379	0.448
							ACC	0.535	0.646	0.585
5	0.017	0.064	50	2000	4.42	0.61	TPR	0.715	0.733	0.721
							FPR	0.543	0.512	0.555
							ACC	0.464	0.494	0.452
6	0.017	0.128	50	4000	5.42	0.50	TPR	0.683	0.642	0.680
							FPR	0.578	0.524	0.574
							ACC	0.425	0.478	0.429

Table 1: Clustering detection rates averaged over 100 simulated point patterns generated on the `chicago` linear network with λ_1 and λ_2 intensities and $\mathbb{E}[n_1]$ and $\mathbb{E}[n_2]$ expected number of points for the first and second point patterns.

4.2 Clustering detection on a linear network without cycles

The linear network `dendrite` is the dendrite network of a neuron (nerve cell) recorded by the Kosik Lab (UCSB) (Jammalamadaka et al., 2013) and available in the package `spatstat.data` (Baddeley et al., 2021b). The `dendrite` network has 640 intersections, 639 uninterrupted segments and a total length of 1934 μm (1.93 mm) (Baddeley et al., 2014), while the sub-network we choose has 245 intersections, 243 uninterrupted segments and a total length of 778 μm (0.78 mm). We generate both point patterns on the network branches, as shown in Figure 5a.

We report the results of the clustering procedure in Table 2. For designs 1 and 2, where $\lambda_1 > \lambda_2$, the clustering method performs well in terms of true-positive rates for $K = 5$, $K = 10$ and $\hat{K}$, but has high false-positive rates. For designs 3 and 4, where $\lambda_1 \approx \lambda_2$, and design 5, where $\lambda_1 < \lambda_2$, the clustering method also performs well in terms of true-positive rates for $K = 5$, $K = 10$ and $\hat{K}$, and has an increase in false-positive rates. For design 6, where $\lambda_1 < \lambda_2$, the clustering method has low performance in terms of true and false positive rates for $K = 5$, $K = 10$ and $\hat{K}$. The true-positive rates are always higher than the false-positive rates. The method performs better when the ratio between the expected number of points of the first and the second point pattern approaches one. This outcome lines up with the previous experiments for networks with loops.



(a) Simulate pattern. (b) Clustering pattern.

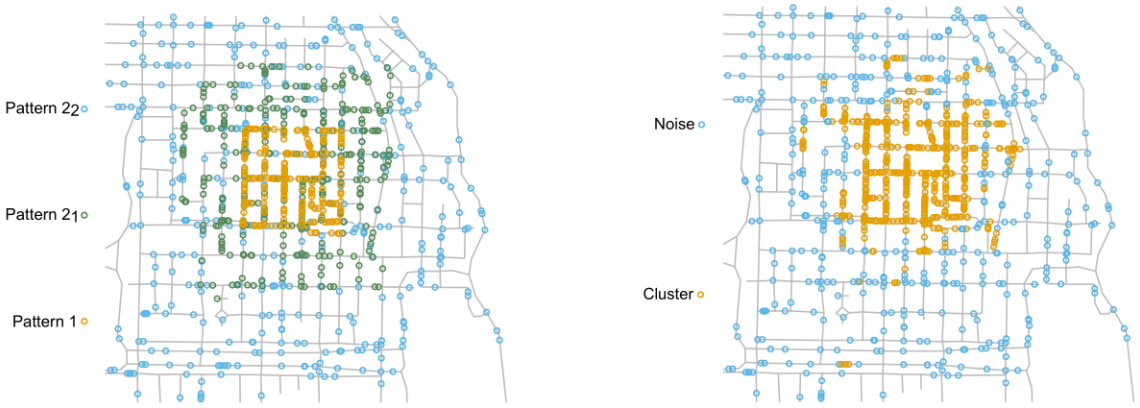
Figure 5: *(a):* Simulated point pattern on the `dendrite` linear network. The first and second point patterns are simulated from a Poisson point process homogeneous with $\lambda_1 = 0.386$, $\mathbb{E}[n_1] = 300$ and $\lambda_2 = 0.207$, $\mathbb{E}[n_2] = 400$, respectively. *(b):* Randomly selected classified point pattern using the proposed clustering method.

								K		
Design	λ_1	λ_2	$\mathbb{E}[n_1]$	$\mathbb{E}[n_2]$	$\bar{K}$	sd(K)	Rate	5	10	$\hat{K}$
1	0.386	0.207	300	400	7.98	0.67	TPR	0.910	0.956	0.941
							FPR	0.478	0.428	0.438
							ACC	0.688	0.737	0.725
2	0.257	0.207	200	400	8.78	0.75	TPR	0.863	0.905	0.894
							FPR	0.504	0.448	0.453
							ACC	0.619	0.670	0.664
3	0.386	0.388	300	750	12.64	1.09	TPR	0.842	0.891	0.910
							FPR	0.532	0.473	0.453
							ACC	0.575	0.631	0.650
4	0.257	0.259	200	500	10.30	1.21	TPR	0.849	0.892	0.892
							FPR	0.539	0.476	0.474
							ACC	0.571	0.628	0.630
5	0.193	0.259	150	500	11.02	1.47	TPR	0.802	0.840	0.846
							FPR	0.541	0.485	0.474
							ACC	0.538	0.590	0.600
6	0.032	0.388	25	750	11.91	1.91	TPR	0.569	0.559	0.548
							FPR	0.560	0.520	0.505
							ACC	0.444	0.483	0.497

Table 2: Clustering detection rates averaged over 100 simulated point patterns generated on the `dendrite` linear network with λ_1 and λ_2 intensities and $\mathbb{E}[n_1]$ and $\mathbb{E}[n_2]$ expected number of points for the first and second point patterns.

4.3 Clustering detection with two nested sub-networks

We consider an even more complex scenario where the points of the second point pattern come from different intensities found in the network. Thus, the cluster looks more confused due to the fact that the points of the first point pattern could be somehow hidden. To show the procedure's usefulness, we simulate from a setting dealing with the mixture of multiple intensities of a homogeneous Poisson process on the observed point pattern. We use again the `chicago` linear network and the sub-network used to generate the first point pattern defined in Section 4.1.



(a) Simulate pattern. (b) Clustering pattern.

Figure 6: *(a):* Simulated point pattern on the `chicago` linear network with two nested sub-networks. The first point pattern is simulated from a homogeneous Poisson point process with $\lambda_1 = 0.067$ and $\mathbb{E}[n_1] = 200$. The second point pattern is simulated from two homogeneous Poisson point processes with $\lambda_{2_1} = 0.026$, $\mathbb{E}[n_{2_1}] = 300$ in the larger sub-network and $\lambda_{2_2} = 0.013$, $\mathbb{E}[n_{2_2}] = 400$ in the whole `chicago` network. *(b):* Randomly selected classified point pattern using the proposed clustering method.

In this case, we consider a third intensity value in a larger nested sub-network than the sub-network used for the first point pattern as shown in Figure 6a; the largest nested sub-network of the `chicago` has 130 intersections, 199 uninterrupted segments and a total length of 11731 feet. We simulate two distinct Poisson point patterns with constant intensities in the two nested sub-networks and superimpose them to a point pattern generated on the whole network. The first point pattern will consist of the points generated in the smaller sub-network. The second point pattern will consist of the points generated in the larger sub-network and those generated in the whole linear network.

Let us assume that the points of the second point pattern generated on the larger sub-network have intensity λ_{2_1} and those generated on the whole linear network have intensity λ_{2_2}. Table 3 shows the results of the clustering procedure of the first and second point patterns for the `chicago` network, where the second point pattern consists of two nested

										K		
Design	λ_1	λ_{2_1}	λ_{2_2}	$\mathbb{E}[n_1]$	$\mathbb{E}[n_{2_1}]$	$\mathbb{E}[n_{2_2}]$	$\bar{K}$	sd(K)	Rate	5	10	$\hat{K}$
1	0.067	0.026	0.013	200	300	400	4.19	0.42	TPR	0.988	0.998	0.980
									FPR	0.423	0.446	0.426
									ACC	0.667	0.651	0.663
2	0.067	0.051	0.032	200	600	1000	4.89	0.31	TPR	0.972	0.994	0.971
									FPR	0.503	0.509	0.505
									ACC	0.551	0.547	0.549
3	0.067	0.085	0.032	200	1000	1000	4.56	0.50	TPR	0.984	0.998	0.981
									FPR	0.622	0.646	0.624
									ACC	0.433	0.413	0.432
4	0.017	0.085	0.032	50	1000	1000	4.37	0.49	TPR	0.942	0.987	0.935
									FPR	0.645	0.673	0.647
									ACC	0.369	0.343	0.368
5	0.033	0.034	0.032	100	400	1000	4.87	0.34	TPR	0.943	0.985	0.943
									FPR	0.531	0.523	0.533
									ACC	0.500	0.510	0.498
6	0.067	0.068	0.064	200	800	2000	5.13	0.34	TPR	0.952	0.977	0.953
									FPR	0.535	0.487	0.532
									ACC	0.499	0.544	0.502
7	0.134	0.034	0.064	400	400	2000	5.03	0.17	TPR	0.975	0.974	0.976
									FPR	0.402	0.281	0.401
									ACC	0.653	0.756	0.654
8	0.033	0.026	0.128	100	300	4000	5.84	0.42	TPR	0.810	0.816	0.815
									FPR	0.587	0.526	0.576
									ACC	0.422	0.482	0.433
9	0.017	0.026	0.064	50	300	2000	4.83	0.47	TPR	0.820	0.847	0.823
									FPR	0.561	0.517	0.567
									ACC	0.447	0.491	0.442
10	0.017	0.051	0.128	50	600	4000	5.97	0.33	TPR	0.811	0.815	0.816
									FPR	0.604	0.544	0.592
									ACC	0.401	0.460	0.412

Table 3: Clustering detection rates averaged over 100 simulated point patterns generated on the `chicago` linear network with two nested sub-networks with λ_1, λ_{2_1} and λ_{2_2} intensities and $\mathbb{E}[n_1]$, $\mathbb{E}[n_{2_1}]$ and $\mathbb{E}[n_{2_2}]$ expected numbers of points for the first point pattern and the two point patterns that compose the second pattern.

intensities. As in the previous settings, the results are computed for different combinations of the intensities λ_1, λ_{2_1} and λ_{2_2}. In designs 1 and 2, we take $\lambda_1 > \lambda_{2_1} > \lambda_{2_2}$. In designs 3 and 4, we take $\lambda_1 < \lambda_{2_1} > \lambda_{2_2}$ for $\lambda_1 > \lambda_{2_2}$ in design 3 and $\lambda_1 < \lambda_{2_2}$ for design 4. In designs 5 and 6, we take $\lambda_1 \approx \lambda_{2_1} \approx \lambda_{2_2}$. In designs 7 and 8, we take $\lambda_1 > \lambda_{2_1} < \lambda_{2_2}$ for $\lambda_1 > \lambda_{2_2}$ in design 7 and with $\lambda_1 < \lambda_{2_2}$ in design 8. In designs 9 and 10, we take $\lambda_1 < \lambda_{2_1} < \lambda_{2_2}$. Looking at the clustering rate in Table 3, we can see a very good method performance for designs 1 to 7 in terms of true positive rates for $\hat{K}$, $K = 5$ and $K = 10$. Designs 8, 9 and 10 decrease the true positive rates, although it is still good. In all designs, we can see different variations of false-positive rates for $\hat{K}$, $K = 5$ and $K = 10$, but consistently lower than true-positive rates. In this scenario, we can also notice that the method performance is better when the ratio between the expected number of points of the first and the second point patterns approaches

one, that is, when the cluster is more intense.

4.4 Clustering detection with two disjoint sub-networks

To highlight the flexibility of the clustering method in terms of the shape of the sub-network where the cluster is generated, we simulate a down-to-earth scenario on the streets of Antonio Narino locality in Bogota, Colombia (Moncada, 2018).

(a) Simulate pattern.

(b) Clustering pattern.

Figure 7: *(a):* Simulated point pattern on the Antonio Narino linear network. The first point pattern is simulated from homogeneous Poisson point processes on two disjoint streets with $\lambda_{1_1} = 0.109$, $\mathbb{E}[n_{1_1}] = 400$ and $\lambda_{1_2} = 0.072$, $\mathbb{E}[n_{1_2}] = 600$. The second point pattern is simulated from a homogeneous Poisson point process with $\lambda_2 = 0.047$ and $\mathbb{E}[n_2] = 6000$. *(b):* Randomly selected classified point pattern using the proposed clustering method.

Antonio Narino is one of the eighteen zones, called Localidades, into which the metropolitan area of Bogota is divided. The road network of Antonio Narino locality has 2062 intersections, 2706 uninterrupted segments and a total length of 128.69 km. The first point pattern are simulated on two of the longest roads, as shown in Figure 7a. The first road (right) is a sub-network with 237 intersections, 254 uninterrupted segments and a total length of 8.32

km and the second road (left) is a sub-network with 106 intersections, 112 uninterrupted segments and a total length of 3.68 km. In the first road, we simulate a pattern with expected number $\mathbb{E}[n_{1_2}]$ and intensity λ_{1_2}. Likewise, in the second road, we simulate another point pattern with expected number $\mathbb{E}[n_{1_1}]$ and intensity λ_{1_1}. In this way, the first point pattern consists of the points generated on both roads, and the second point pattern consists of the points generated in the whole Antonio Narino linear network. In this scenario, there are two independent point clusters.

										K		
Design	λ_{1_1}	λ_{1_2}	λ_2	$\mathbb{E}[n_{1_1}]$	$\mathbb{E}[n_{1_2}]$	$\mathbb{E}[n_2]$	$\bar{K}$	sd(K)	Rate	5	10	$\hat{K}$
1	0.109	0.072	0.047	400	600	6000	5.07	0.26	TPR	0.966	0.904	0.967
									FPR	0.432	0.189	0.430
									ACC	0.625	0.825	0.627
2	0.054	0.036	0.023	200	300	3000	5.58	0.65	TPR	0.969	0.924	0.962
									FPR	0.445	0.224	0.418
									ACC	0.615	0.798	0.637
3	0.054	0.072	0.047	200	600	6000	5.84	0.56	TPR	0.956	0.903	0.954
									FPR	0.481	0.259	0.450
									ACC	0.570	0.760	0.597
4	0.014	0.072	0.047	50	600	6000	6.63	0.76	TPR	0.949	0.886	0.948
									FPR	0.492	0.239	0.447
									ACC	0.551	0.773	0.592
5	0.094	0.093	0.093	345	775	12000	6.32	0.63	TPR	0.935	0.928	0.942
									FPR	0.539	0.395	0.510
									ACC	0.501	0.633	0.529
6	0.034	0.036	0.035	125	300	4500	7.29	0.89	TPR	0.922	0.940	0.937
									FPR	0.513	0.427	0.479
									ACC	0.525	0.604	0.557
7	0.094	0.024	0.047	345	200	6000	8.02	0.64	TPR	0.913	0.776	0.853
									FPR	0.503	0.218	0.356
									ACC	0.531	0.781	0.661
8	0.034	0.024	0.047	125	200	6000	5.18	0.73	TPR	0.839	0.880	0.841
									FPR	0.540	0.497	0.538
									ACC	0.479	0.522	0.481
9	0.014	0.024	0.047	50	200	6000	4.92	0.61	TPR	0.807	0.855	0.806
									FPR	0.546	0.514	0.547
									ACC	0.468	0.501	0.467
10	0.014	0.024	0.093	50	200	12000	4.98	0.14	TPR	0.713	0.745	0.713
									FPR	0.572	0.527	0.573
									ACC	0.434	0.478	0.433

Table 4: Clustering detection rates averaged over 100 simulated point patterns generated on the Antonio Nariño linear network with λ_{1_1}, λ_{1_2} and λ_2 intensities and $\mathbb{E}[n_{1_1}]$, $\mathbb{E}[n_{1_2}]$ and $\mathbb{E}[n_2]$ expected number of points on the first road, second road and the whole linear network.

The results of the clustering procedure is shown in Table 4 for different combinations of intensities λ_{1_1}, λ_{1_2} and λ_2. In designs 1 and 2, we take $\lambda_{1_1} > \lambda_{1_2} > \lambda_2$. In designs 3 and 4, we take $\lambda_{1_1} < \lambda_{1_2} > \lambda_2$ with $\lambda_{1_1} > \lambda_2$ in design 3 and $\lambda_{1_1} < \lambda_2$ in design 4. In designs 5

and 6, we take $\lambda_{1_1} \approx \lambda_{1_2} \approx \lambda_2$. In designs 7 and 8, we take $\lambda_{1_1} > \lambda_{1_2} < \lambda_2$ with $\lambda_{1_1} > \lambda_2$ in design 7 and $\lambda_{1_1} < \lambda_2$ in design 8. In designs 9 and 10, we take $\lambda_{1_1} < \lambda_{1_2} < \lambda_2$. The clustering detection rates of the procedure in Table 4 have a similar interpretation to those in the previous scenario in Section 4.3, showing a good clustering method performance in all designs from 1 to 10.

5 Identifying high- and low-density road traffic accident in two cities of Colombia

A traffic accident is an involuntary event generated by at least one vehicle in motion on roads in which there is damage to vehicles and objects and injuries or death to the people involved (WHO, 2018). According to WHO (2018), traffic accidents are the eighth cause of death worldwide and the leading cause of death for people between 5 and 29 years of age. Particularly in Colombia, ITF (2019) reports that between 2010 and 2018, the number of dead people in road accidents has been upward, and the annual number of these deaths increased by nearly 30%. Based on the report of the National Road Safety Agency of Colombia ANSV (2021), from 2016 to 2019, the number of dead people in road accidents annually remained between 3.4% and 3.7% of the total traffic accidents. When a traffic accident occurs, a chain of actions of different private and public entities begins to clarify what happened and determine the cause of this traffic accident. This means that the investigation of the reasons for which the accident is generated begins after its occurrence. Some factors that generally produce a traffic accident are human, environmental or mechanical. These factors change according to the country or city where the accidents occur; therefore, traffic accidents can be seen as events randomly distributed on a road network, and since their study is carried out once the event has occurred, we can consider that an underlying point process governs them on the linear network.

We are interested in finding the set of streets where serious traffic accidents are most likely to occur in a network of roads in a city. The collection of such streets must be formed mainly by a point process of high density that has a different intensity from the set of locations where serious traffic accidents are less likely to occur and which should occur over the whole street network of the city. The point pattern that occurs on the complete street network has a low density and makes sense in the context of severe traffic accidents to the extent that anywhere in the city, there may be streets that have a lower number of severe traffic accidents due to low traffic, impediments to high speeds, or other factors that decrease the risk of fatalities.

5.1 Road traffic injuries and deaths in Bogota City

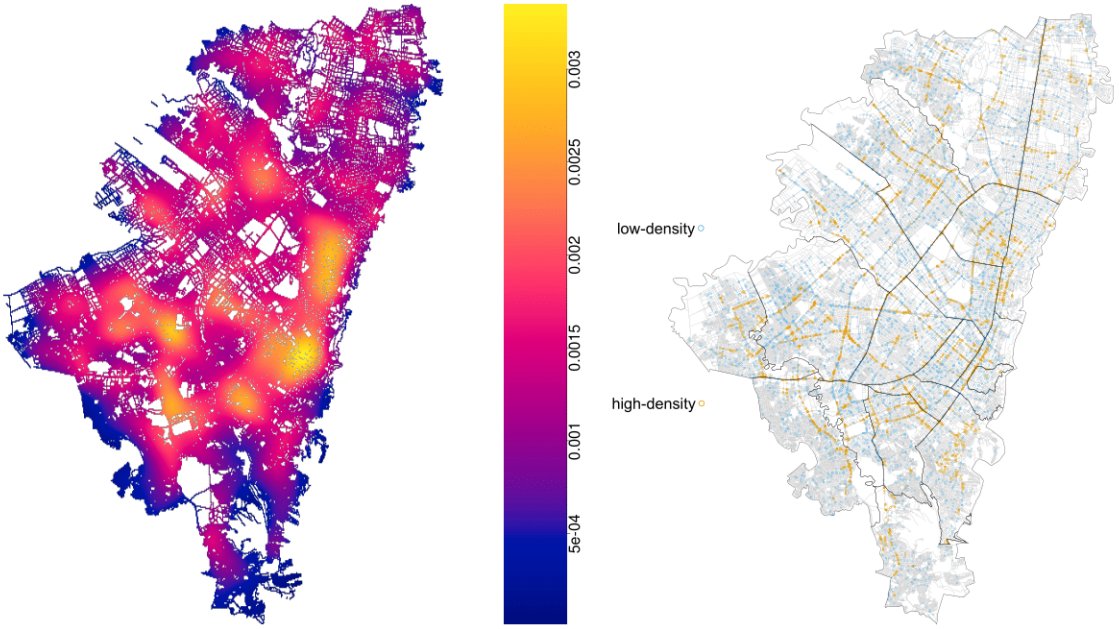
Bogota is Colombia's capital and the largest and most populous city. In 2015, this city was estimated to have 2148541 vehicles (Secretaría Distrital de Movilidad, 2015). The simplified city road network used in this work consists of 199135 intersections, 243157 uninterrupted

segments and a total length of 8218.90 km. We analyzed a database of 10979 traffic accidents that resulted in injuries or deaths in 2015 in the urban area of the city, published in the OpenData portal of Bogota Town Hall, together with the road network shapefile (Moncada, 2018; Secretaría Distrital de Movilidad, 2022). The city of Bogota is divided into eighteen zones called localities. Identifying roads with high-density traffic injuries and deaths provides valuable information for decision-making in public policies that help prevent more of these unfortunate events. In the context of traffic accidents, we aim to employ our developed methodology in linear networks that allow us to distinguish highly congested zones in the sense of traffic accidents from those with lower intensity. The dataset consists entirely of traffic accidents and thus the classification of the traffic accidents into two distinct groups based on intensity: the most accidents rate and less accidents rate zones. By applying this innovative approach to separate patterns with higher traffic accident density from those with lower density, we can effectively identify and prioritize areas requiring additional attention and intervention regarding accident prevention and safety measures.

The application of the proposed clustering procedure to analyze traffic accidents in Bogota directly using the whole network has a high computational cost, and the procedure does not yield results due to its size and complexity. To apply the proposed clustering process to the Bogota network, it must be assumed that both overlapping point patterns are Poisson processes, which implies that the occurrence of events is independent in the disjoint sub-networks of this network. Therefore, we applied the clustering procedure independently in each of the sub-networks of the eighteen localities of Bogota; then we consider the union of all these point patterns classified as the point pattern classified of traffic accidents in the whole network of Bogota as shown in Figure 8b.

We estimate the non-parametric kernel-based intensity using a two-dimensional Gaussian kernel with a bandwidth of $\sigma = 600$ meters (Rakshit et al., 2019b) shown in Figure 8a. The bandwidth used was found empirically, starting from Scott's rule as a reference (Scott, 2015) and then increasing it accordingly so that the hot spots would be displayed on the intensity plot. The kernel estimate of accident intensity in Bogota shows that most traffic accidents occur in different sectors around the city centre. However, a large part of the road network has intermediate intensity values, making it difficult to decide whether it is a road with a high accident rate. In this way, it would be possible to delimit some sub-networks where the probability of occurrence of new traffic accidents is high, but the roads with intermediate values make it difficult to decide on the delimitation of the sub-networks that must be intervened to avoid new traffic accidents.

The clustering of the traffic accidents in the city of Bogota during 2015 shown in Figure 8b is depicted as a multitype point pattern for a dichotomous mark, in which it is possible to identify the road segments with a high density of traffic injuries and deaths in the city of Bogota. Therefore, the classification procedure can help to decide whether the road segments with intermediate intensity values are significantly relevant to delimit the sub-network and use this as a companion tool of the intensity to study the behaviour of traffic accidents in Bogota.



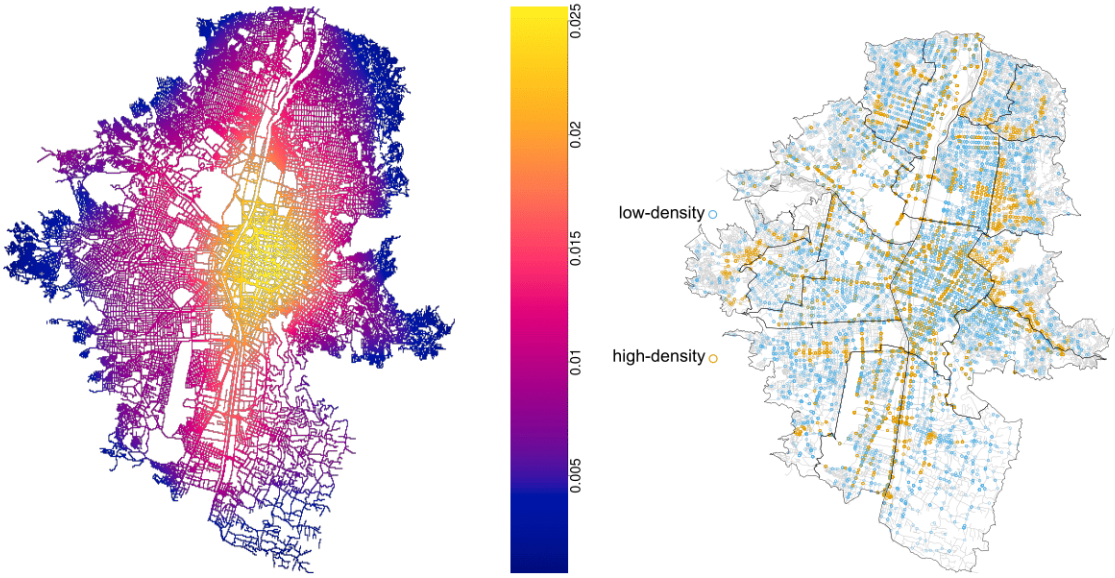
(a) Estimated intensity. (b) Clusters of the traffic accidents.

Figure 8: Traffic accidents on the road network of Bogota in which at least one person was injured or death recorded in the year 2015. *(a):* Estimated non-parametric kernel-based intensity with bandwidth $\sigma = 600$ meters. *(b):* Clusters of the traffic accidents in the localities of Bogota.

In detail, the estimated intensity of the traffic injuries and death accidents in Bogota has a couple of large hot spots in the central-eastern zone of the city. This zone includes the downtown, the Palace of Justice and the national capitol, among other important places, surrounded by government offices, museums, churches and some historic buildings where many vehicles generally move. Additionally, an important avenue crosses the entire city from south to north and has access to many other avenues that pass through this zone. Other relevant hot spots in the city are located in the north zone, where the city's financial, cultural, and recreational activity places are concentrated. The centre-west is another zone with a high probability of traffic accidents where large factories, parks, sports centres, administrative offices, and the international airport are located, which means that the roads that cross this zone are generally crowded with vehicles. Based on the estimated intensity and considering the clustering results, it is possible to appreciate more explicitness of the road segments with high-density accidents in the hot spot zone. In summary, while Figure 8a shows the zones with the highest intensity values, the output in Figure 8b helps to the delimitation of the sub-networks of higher traffic accidents and, in this way, to define which roads must be intervened with the highest priority to prevent traffic accidents.

5.2 Road traffic injuries and deaths in Medellin city

Medellin is the second most important city in Colombia after Bogota. In 2019 it was estimated that this city had 1756893 vehicles (Medellín cómo vamos, 2021). The city's road network consider in this study consists of 110533 intersections, 115939 uninterrupted segments and a total length of 2306.48 km. We analysed a database of 22743 traffic accidents that resulted in injuries or deaths in 2019 in the city, also published in the OpenData portal of Medellin Town Hall, and the road network shapefile (Secretaría de Movilidad, 2022). The city of Medellin is divided into sixteen zones called communes. Analogously as we did with Bogota, in Figure 9b, we can see that we apply the clustering procedure within each of the communes and then unify the classified point patterns to form the clustering point pattern in the whole city of Medellin.



(a) Estimated intensity. (b) Clusters of the traffic accidents.

Figure 9: Traffic accidents on the road network of Medellin in which at least one person was injured or death recorded in the year 2019. *(a):* Estimated non-parametric kernel-based intensity with bandwidth $\sigma = 700$ meters. *(b):* Clusters of the traffic accidents in the communes of Medellin.

Figure 9a shows the non-parametric kernel-based intensity estimated with bandwidth $\sigma = 700$ meters. The intensity of the traffic injuries and death accidents in Medellin has a single large hot spot in the central-eastern zone, which is the commercial and financial centre of the city. This zone includes the downtown, a large part of the city's commerce, residences, factories, administrative offices, museums, churches, and historic buildings. Additionally, four important avenues cross the hot spot, and many vehicles circulate daily in the city using these roads. Based on the intensity in Figure 9a and the result of the clustering procedure of

the traffic injuries and death accidents in Medellin in 2019 shown in Figure 9b, it is possible to differentiate between the road segments with the highest number of accidents in the hot spots. Therefore, the combination of the information provided by these two exploratory tools makes it more feasible to interpret the results in geometric spaces that are not intuitive.

6 Conclusions

In this paper, we explore point process clustering detection on linear networks by focusing on the classification of two distinct types of point patterns observed in a point pattern along the network. Our novel clustering method is designed to effectively discern between these point patterns and leverages parameter estimations derived from an EM (Expectation-Maximization) algorithm. A notable feature of our approach is its ability to automatically determine the optimal number of nearest neighbours to consider in the statistical analysis, achieved through a segmented regression model. Our methodology is versatile enough to operate without further assumptions regarding the shapes of the point patterns, and it is grounded in the assumption that both superimposed point patterns come from Poisson point processes.

The simulation study highlights the performance and accuracy of the proposed procedure in terms of clustering detection rates. The clustering procedure applied to the real point pattern dataset of traffic accidents occurring on a city's road network, together with the estimation of intensity, can be a helpful tool in decision-making when delimiting the hot-spot roads that have priority to be intervened to prevent these accidents.

Given the growing interest in linear network data, we believe the proposed approach could be used in many application contexts. Future work may consider modifying the EM algorithm to estimate the intensities of more than two groups of patterns simulated over the same linear network or extending this methodology to the spatio-temporal case on linear networks.

Acknowledgments

The research work of Juan F. Díaz-Sepúlveda and Francisco J. Rodrǵuez-Cortés has been partially supported by Universidad Nacional de Colombia, HERMES projects, Grant/Award Number: 56470. The research work of Giada Adelfio and Nicoletta D'Angelo has been supported by the Targeted Research Funds 2023 (FFR 2023) of the University of Palermo (Italy) and by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C93C22005270001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

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

Ang, Q. W., Baddeley, A., and Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics*, 39(4):591–617.

ANSV (2021). *Informe Sobre Seguridad Vial Para el Congreso de la República de Colombia*. Agencia Nacional de Seguridad Vial, Bogota.

Baddeley, A., Jammalamadaka, A., and Nair, G. (2014). Multitype point process analysis of spines on the dendrite network of a neuron. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(5):673–694.

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., and Davies, T. M. (2021a). Analysing point patterns on networks — a review. *Spatial Statistics*, 42:100435. Towards Spatial Data Science.

Baddeley, A., Turner, R., and Rubak, E. (2021b). *spatstat.data: Datasets for 'spatstat' Family*. R package version 2.1-2.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

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.

Cronie, O., Moradi, M., and Mateu, J. (2020). Inhomogeneous higher-order summary statistics for point processes on linear networks. *Statistics and computing*, 30(5):1221–1239.

D'Angelo, N. and Adelfio, G. (2023). Selecting the kth nearest-neighbour for clutter removal in spatial point processes through segmented regression models. In *Book of Abstracts GRASPA 2023, ISBN 979-12-210-3389-2*.

D'Angelo, N., Adelfio, G., and Mateu, J. (2021). Assessing local differences between the spatio-temporal second-order structure of two point patterns occurring on the same linear network. *Spatial Statistics*, 45:100–534.

D'Angelo, N., Adelfio, G., and Mateu, J. (2023). Local inhomogeneous second-order characteristics for spatio-temporal point processes occurring on linear networks. *Statistical Papers*, 64(3):779–805.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Federer, H. (1969). *Geometric measure theory*. Springer, Heidelberg.

González, J. A., Rodríguez-Cortés, F. J., Romano, E., and Mateu, J. (2021). Classification of events using local pair correlation functions for spatial point patterns. *Journal of Agricultural, Biological and Environmental Statistics*, 26(4):538–559.

ITF (2019). *Road Safety Annual Report 2019*. OECD Publishing, Paris.

Jammalamadaka, A., Banerjee, S., Manjunath, B. S., and Kosik, K. S. (2013). Statistical analysis of dendritic spine distributions in rat hippocampal cultures. *BMC bioinformatics*, 14(1):1–19.

Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(3):231–240.

McSwiggan, G., Baddeley, A., and Nair, G. (2017). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics*, 44(2):324–345.

McSwiggan, G., Baddeley, A., and Nair, G. (2020). Estimation of relative risk for events on a linear network. *Statistics and Computing*, 30.

Medellín cómo vamos (2021). *Informe de Calidad de Vida de Medellín, 2020*. Medellín cómo vamos, Medellín.

Moncada, I. (2018). Análisis espacio-temporal de los accidentes de tránsito en bogotá utilizando patrones puntuales. *Tesis de Maestría. Universidad Nacional de Colombia*.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill.

Moradi, M. M., Rodríguez-Cortés, F. J., and Mateu, J. (2018). On kernel-based intensity estimation of spatial point patterns on linear networks. *Journal of Computational and Graphical Statistics*, 27(2):302–311.

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

Rakshit, S., Baddeley, A., and Nair, G. (2019a). Efficient code for second order analysis of events on a linear network. *Journal of Statistical Software*, 90(1):1–37.

Rakshit, S., Davies, T., Moradi, M. M., McSwiggan, G., Nair, G., Mateu, J., and Baddeley, A. (2019b). Fast kernel smoothing of point patterns on a large network using two-dimensional convolution. *International Statistical Review*, 87(3):531–556.

Rakshit, S., McSwiggan, G., Nair, G., and Baddeley, A. (2021). Variable selection using penalised likelihoods for point patterns on a linear network. *Australian & New Zealand Journal of Statistics*, 64(1):1 – 38.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, New York, 2 edition.

Secretaría de Movilidad (2022). *GeoMedellín*. Alcaldía de Medellín.

Secretaría Distrital de Movilidad (2015). *Movilidad en cifras 2015*. Alcaldia Mayor de Bogotá, Bogotá.

Secretaría Distrital de Movilidad (2022). *Sistema Integrado de Información sobre Movilidad Urbana Regional*. Alcaldía de Bogotá.

WHO (2018). *Global status report on road safety 2018*. World Health Organization, Geneva.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2:165–193.