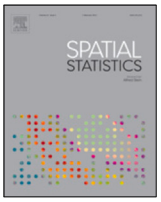
Spatial Statistics 51 (2022) 100693

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Linear hotspot detection for a point pattern in the vicinity of a linear network

Jacob Modiba [a], Inger Fabris-Rotelli [a,*], Alfred Stein [b,a], Gregory Breetzke [c]

[a] Department of Statistics, University of Pretoria, South Africa
[b] Department of Earth Observation Science, University of Twente, Netherlands
[c] Department of Geography, Geoinformatics and Meteorology, University of Pretoria, South Africa

ARTICLE INFO

Article history:
Received 3 March 2022
Received in revised form 4 August 2022
Accepted 9 August 2022
Available online 29 August 2022

Keywords:
Linear network
Point pattern
Crime analysis
Linear connectivity
Khayelitsha

ABSTRACT

The analysis of point patterns on linear networks is receiving current attention in spatial statistics. This refers to the analysis of points in a spatial domain that coincide with a linear network like a road network. The linear network is modelled as a set of lines that are connected at their ends or are intersecting, that is, modelled as mathematical graphs. Limited research so far has been conducted on spatial points that fall on the Euclidean space containing the linear network. This study addresses new steps by exploring points in the vicinity of the network that do not necessarily fall on the linear network. We present a novel method that is motivated by crime locations amongst a road network. The aim is to detect spatial hotspots around a linear network, where crime locations are considered as a point pattern lying in the vicinity of the linear road network. A new connectivity measure is also introduced to define the line segment neighbours of a line segment. The methodology is applied to crime data in Khayelitsha, South Africa. We detect a pattern of crime locations within the network that can be well interpreted. We conclude that our method is well applicable and could potentially help governmental organisations to allocate measures to reduce criminality.

* Corresponding author.
E-mail address: inger.fabris-rotelli@up.ac.za (I. Fabris-Rotelli).

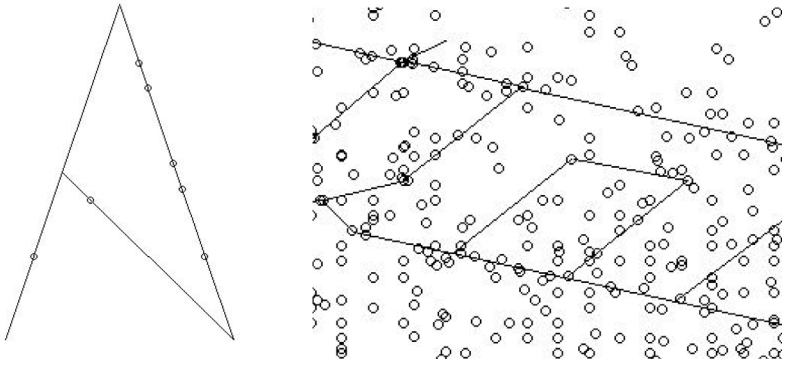
https://doi.org/10.1016/j.spasta.2022.100693
2211-6753/

1. Introduction

A linear network is a union of a finite collection of lines in a plane (Ang et al., 2012). There are two types of network constrained point data. The most common type is points that lie directly on the lines of a linear network (Fig. 1(a)) and where the shortest path on the network, namely a network distance, is used to calculate the distance. Some examples of events restricted to linear networks are the locations of trees along the streets or rivers (Spooner et al., 2004), road accidents and car-jacking on street networks (Yamada and Thill, 2003). The other type is points lying in the vicinity of the linear network (Fig. 1(b)), namely within the Euclidean space in which the linear network is embedded. An example are crimes that occur in the vicinity of the road network e.g. mugging.

There has been a recent increase in spatial statistics methodology for spatial linear networks. The focus there has been for points lying on a linear network, such as, density estimation of points on a network (Borruso, 2005, 2008; Moradi et al., 2018; Mateu et al., 2020), distance metrics and second order analysis (Rakshit et al., 2017), space–time analysis of points on a network (Eckardt and Mateu, 2016), regression for points on a network (Eckardt and Mateu, 2018), summary statistics (Cronie et al., 2020), directed networks (Rasmussen and Christensen, 2021), the use of network based quadrants to calculate the features (Shino, 2008), the moving-segment approach, which uses distances along the network and also includes the connectivity at intersections of networks (Steenberghen et al., 2010), and local indicators of network-constrained clusters (LINCS) (Yamada and Thill, 2007), spatio-temporal analysis of points lying on a network (D'Angelo et al., 2021) and a review (Baddeley et al., 2021). The review (Baddeley et al., 2021) highlights important issues when dealing with a network space. The first is the intrinsic lack of homogeneity of the lines in the network, with homogeneity as a common assumption in spatial analysis. Any spatial analysis conducted should therefore account for this with a suitable distance metric. The theory for points that lie directly on a linear network has been developed and implemented in literature. However, to our knowledge, Comas et al. (2019) is the only existing research on points within the vicinity of a linear network. Comas et al. (2019) develop methodology for the correlation between the points and the network. This is an important first step to confirm that in fact there is a relationship between the linear network and the spatial point pattern data that is worth modelling. We propose new methodology for the analysis of a point pattern in the vicinity of a linear network. This approach effectively uses a linear network as an additional spatial structure in a spatial point pattern.

A spatial analysis on linear networks refers to spatial statistical methodology used to analyse point events that occur along linear networks (Okabe and Sugihara, 2012). If the points are in



(a) A point pattern observed on a linear network

(b) A point pattern observed in the vicinity of a linear network

Fig. 1. Possible point patterns in a linear network space.

the vicinity of a linear network, however, traditional spatial analysis needs to differ from current network spatial analysis since it assumes a continuous Euclidean space as opposed to the network constrained space (Jiang and Okabe, 2014). The methodology of this paper provides a mechanism for detection of hotspots via analysis of the determined density for points that lie in the vicinity of a linear network.

A hotspot is defined as a sub-area within a larger area that has a higher than average number of points as compared to the neighbouring areas (Chakravorty, 1995). Local indicators of spatial association can be used to identify statistically significant hotspots. These include the local Getis–Ord (G^*) statistic (Getis and Ord, 1995), Anselin's local Moran's I (Anselin, 1995) and the local Geary's C (Anselin, 1995, 2019). These techniques have been implemented in various studies, for instance the identification of hot- and coldspots of orchard trees using the Getis–Ord (G^*) statistic (Peeters et al., 2015) and finding clusters of high impact accidents (Songchitruksa and Zeng, 2010). Local indicators rely on an initial partition of the spatial area into non-overlapping sub-areas most often using an overlaid grid of a chosen resolution. The number of points are then obtained in each sub-area and used as features where counts in neighbouring areas, i.e., touching grid cells, serve as spatial weights. Each cell is then determined as a significant hot or cold spot. To extend methodology for points within the vicinity of a linear network, we need a mechanism to determine line segments, the equivalent of sub-areas, and a definition of the connectivity for these line segments within the linear network. We make use of the ideas in Tompson et al. (2009) to determine line segments, and define a new way to measure connectivity of these line segments. The possibility of such line segments as crime facilitators is considered here and is covered in more detail in Section 3.

In this paper, we propose a methodology for statistically analysing a spatio-temporal point pattern located in the vicinity of the lines of a linear network. The focus is on determining the implied density of the points onto the linear network and on detecting statistically significant linear hotspots. We also introduce and define the concept of a linear hotspot. This methodology is compared with traditional hotspot detection, namely hotspot analysis on a grid of the point pattern only, ignoring the effect of an associated linear network. The study is inspired by crime data from a large township in South Africa.

The paper is organised as follows. Section 2 gives an overview of the existing hot route methodology of Tompson et al. (2009) and our proposed linear hotspot methodology. Section 3 provides the result when the new methodology is applied on crime data from Khayelitsha, South Africa. The results are discussed in Section 4. And finally, the study is concluded in Section 5.

2. Methodology

We consider a linear network and a related point pattern. A point process is a random mechanism whose outcomes are point patterns (Diggle, 2013). A point pattern is a finite unordered set $x = \{x_1, x_2, \ldots, x_{n_p}\}$ resulting from a point process, where each point x_i represents a location. Let $\ell = \{tv_1 + (1-t)v_2 : 0 \leq t \leq 1\}$ represent a line segment with endpoints v_1 and v_2. The union of a number of such line segments, $L = \bigcup_{i=1}^{n} \ell_i$, is called a linear network, namely a finite collection of line segments $\{\ell_1, \ldots, \ell_n\}$ in the Euclidean plane (Ang et al., 2012). A point pattern in the vicinity of a linear network is a finite unordered set $x = \{x_1, x_2, \ldots, x_{n_p}\}$, where each point x_i represents a location in the vicinity of a linear network (D'Angelo et al., 2021)

We define a *linear hotspot* as a line segment that exhibits a higher than average number of spatial points associated with it than its neighbouring lines segment. Note that this definition holds for both points lying on a linear network as well as points within a vicinity of a linear network. We focus on the latter. In order to detect linear hotspots, an appropriate unbiased division of the network space into line segments is required, mimicking the idea of grid cells in traditional hotspot detection. Furthermore, a definition of which line segments can be considered neighbours is needed. Grid cells are considered neighbours using 4- or 8-connectivity which is an obvious choice for a spatial analysis in a Euclidean space but not for spatial analysis in the vicinity of a linear network. Due to the structure of a linear network, the choice of the neighbourhood structure for spatial analysis in the vicinity of a linear network requires a deeper insight since this has not been done before.

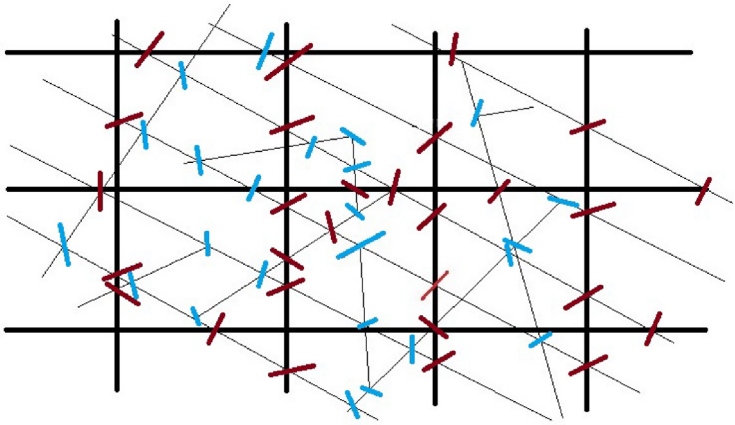


Fig. 2. Line segment determination using the grid cell intercept locations with the linear network (red lines) and the linear network intercept points within a grid cell (blue lines) as introduced in Tompson et al. (2009). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.1. Background methodology

A local intensity measure (Diggle, 2013) is required to detect linear hot- and coldspots. The intensity of a point pattern is the expected number of points per unit area, and is defined (Diggle, 2013) as

$$\lambda(x) = \lim_{n\to 0}\left(\frac{E[N(dx)]}{|dx|}\right)$$

where $E[N(dx)]$ is the expected number of points in a dx and $|dx|$ is the area of the region. Determining the local intensity in a linear network space involves splitting the linear network into smaller line segments. The method introduced in this paper extends the hot route methodology of Tompson et al. (2009). The extension is to points that are in the vicinity of a linear network instead of points that lie directly on the linear network and importantly, to identify statistically significant linear hotspots. Traditional hotspot detection relies on a grid overlaid on the spatial window.

The hot route methodology (Tompson et al., 2009) measures the distribution of points on a linear network by obtaining the rate of events per line segment. The methodology does not conduct linear hotspot detection, but calculates and visualises a density on each line segment. The significance of the densities with respect to neighbouring line segments is not investigated in Tompson et al. (2009). We add the methodology here for completeness, and provide an illustration in Fig. 2

1. Create a grid overlaid on the spatial area of interest and linear network L (bold lines in Fig. 2).
2. Split every line ℓ_i where it intersects the grid cell boundaries (red lines in Fig. 2).
3. Split every line ℓ_i at each line intersection (blue lines in Fig. 2).
4. Relabel the lines as $\{\ell_{ik}\}$ for $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, k_i$, n the number of lines in a linear network and k_i is the number of splits for each original line ℓ_i.
5. Calculate the length $|\ell_{ik}|$ of each line segment.
6. Count the number of points n_{ik} lying on each line segment.
7. Determine the rate of points per line segment as the number of points divided by the length of the line they occur on, $r_{ik} = \frac{n_{ik}}{|\ell_{ik}|}$.
8. Visualise the rate of event per line segment for each line segment

The values r_{ik} are visualised on the linear network as a density (Tompson et al., 2009).

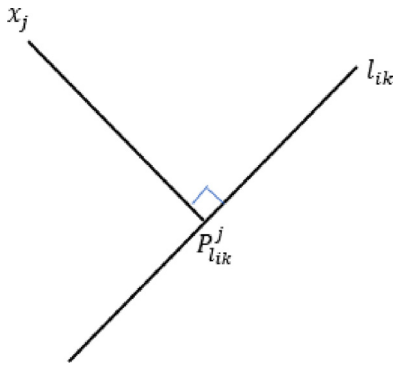


Fig. 3. Illustration of the unique perpendicular distance of point x_j to line segment ℓ_{ik}.

How to choose the grid size is an important consideration. There is no unique way to define a grid and therefore no unique way to split the linear network into line segments. Even a same size grid can be positioned differently. The grid should therefore be chosen in a data-driven way. We discuss this with the crime application later.

2.2. Proposed methodology

The hot route methodology of Tompson et al. (2009) was designed for points that lie directly on a linear network, and points are then uniquely allocated to a line segment. In reality, points do not always lie directly on a linear network, but rather in the vicinity of the linear network. We extend the hot route methodology of Tompson et al. (2009) to accommodate such a case.

The first step is to randomly split the linear network into line segments, as discussed in the hot route methodology. The choice of the grid cells does have an effect on the size of the line segments, since small grid cells will mean smaller line segments. The higher the resolution of the grid however, the larger the increase in the complexity of the computations. Further however, if the resolution is too high, the line segments will be too small to be informative. The same is true if the line segments are too large — in which case a long line segment could be classified as a hotspot with points only present in a portion of the line segment. The choice of the grid cell also depends on the overall number of points in the point pattern and their extent of clustering, since smaller grid cells might result in empty cells with respect to the point pattern. This choice of grid size holds for many approaches in spatial analysis and GIS, and should be chosen appropriately based on the data under consideration, i.e. in a data driven manner. There is thus a decision to be made between grid size choice and computations. There is growing interest in assessing crime at very fine granularities, such as street segments, for a variety of reasons, one of which being that coarse granularities obscure crime hotspots, but counts may become unreliable and unrepresentative if granularities are too fine. Therefore there should be a balance between the two. See Ramos et al. (2021) for an investigation of the effect of the grid size.

Once the grid choice has been made, the points of the point pattern can be assigned to a line segment in one of two ways. A point is allocated to only one line segment by projecting the point to its nearest line segment. Let n_s be the total number of line segments ℓ_{ik} obtained from splitting the lines in a linear network and let the point pattern be $x = \{x_1, x_2, \ldots, x_m\}$. Each point x_j is assigned to line segment ℓ_{ik} if the Euclidean distance $\|x_j - p^j_{\ell_{ik}}\|$ is a minimum, where $p^j_{\ell_{ik}}$ is the position on ℓ_{ik} closest to x_j measured perpendicular from the point to the line segment. This is visualised in Fig. 3.

The counts for each line segment ℓ_{ik} are then calculated as,

$$c_{ik} = \sum_{j=1}^{m} 1^j_{\ell_{ik}} \tag{1}$$

where

$$1^j_{\ell_{ik}} = \begin{cases} 1 & \text{if } \|p^j_{\ell_{ik}} - x_j\| \text{ is a minimum} \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

Alternatively, in a less deterministic mechanism, a point can be allocated to more than one line segment by using weights according to how far they are from each line segment. The counts are then computed as,

$$c^w_{ik}(r) = \sum_{j=1}^{m} w^j_{\ell_{ik}} 1^j_{\ell_{ik}}(r) \tag{2}$$

where

$$1^j_{\ell_{ik}}(r) = \begin{cases} 1 & \text{if } \|p^j_{\ell_{ik}} - x_j\| \leq r \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Here $w^j_{\ell_{ik}}$ is the weight between point x_j and line segment ℓ_{ik} calculated as the inverse distance (Suryowati et al., 2018) between a point and a line segment

$$w_{ij} = \frac{1}{\|p^j_{\ell_{ik}} - x_j\|}.$$

There are different ways to choose the weight matrix, such as the inverse exponential function $w_{ij} = e^{-\|p^j_{\ell_{ik}} - x_j\|}$. Ultimately, the aim is to apply more weight between a point and a line segment if the point is closer to that line segment than to any other line segment. There is no unique way of assigning weights, but for spatial analysis the simplest and most commonly used weight is the inverse distance. Measures other than the distance can also be used to assign weights. Consider for example a method including covariates (Ejigu and Wencheko, 2020). Furthermore, due to spatial dependence reducing as distance increases, only points that are within a certain radius r of the line segments of interest are allocated the weights. The new methodology uses a weighted count of points for each line segment such that points are given weights depending on how far they are from the line segment i.e., small weight for points that are far. In our application there are some points that are far away from the linear network. We do no not need to remove the points since the method is designed to give small weight to the point that are far, therefore they will be automatically given a very small weight or excluded completely based on the radius chosen. This is visualised in Fig. 4.

To obtain the rate of events per unit distance for each line segment ℓ_{ik}, the counts are divided by the corresponding standardised length of the line segment,

$$r_{ik} = \frac{c_{ik}}{|\ell_{ik}|} \text{ or } \frac{c^w_{ik}}{|\ell_{ik}|}. \tag{3}$$

The r_{ik}'s are used to visualise the local density of points for each line segment.

2.2.1. Linear neighbours

Having calculated the rates per line segment, we next identify statistically significant line segments which we call linear hotspots (coldspots). This requires one to identify the line segments with higher (lower) than average number of points in their vicinity as compared to its neighbouring line segments. This necessitates a new definition of neighbours for line segments.

Fig. 5 represents the neighbourhood structure that is introduced with this methodology. The blue line segments represent the line segment of interest and the black line segment represent its neighbours. The neighbourhoods are defined as line segments that share an edge (start or end point), represented in Fig. 5(a); line segments that are within a certain radius from start and end point of the line segment of interest, represented in Fig. 5(b); and line segments that are within a certain radius from the midpoint of the line segment of interest, represented in Fig. 5(c). The three

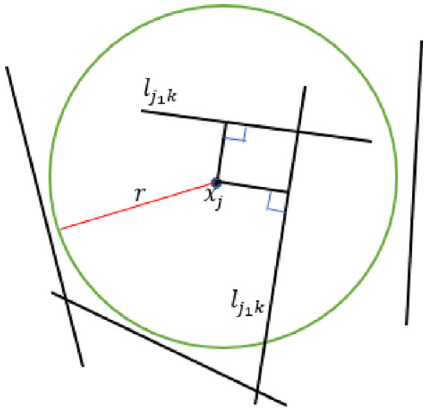


Fig. 4. Illustration of the weighted distances from point x_j to the line segments with a radius r.

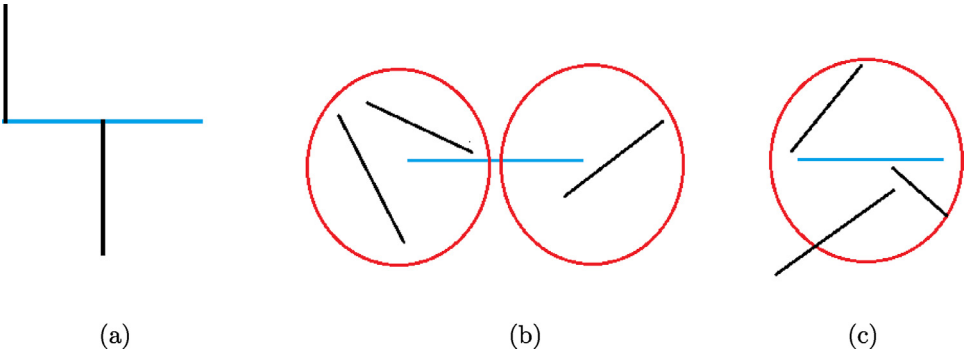


Fig. 5. The neighbourhood structures: (a) Line segments that share an edge (black) with the line segment of interest (blue), (b) Line segments (black) that are within a certain radius away from the start and end point of the line segment of interest (blue), (c) Line segments that are within a certain radius of the midpoint of the line segment of interest (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

neighbour definitions can be considered separately or as a combination. We define this formally in Definition 1.

A weight matrix $E = [e_{st}]$ is used to represent the structure of the line segment neighbours where line segments ℓ_s and ℓ_t are labelled neighbours if $e_{st} = 1$. Here s and t are indexes in $I = \{ik : i = 1, 2, \ldots, n, k = 1, 2, \ldots, k_i\}$.

Definition 1. Let ℓ_{k_1} and ℓ_{k_2} be two line segments represented as

$$\ell_{k_1} = \{p_{k_1} = (x_{k_1}, y_{k_1}) : y_{k_1} = m_{k_1}x_{k_1} + c_{k_1}\} \tag{4}$$

and

$$\ell_{k_2} = \{p_{k_2} = (x_{k_2}, y_{k_2}) : y_{k_2} = m_{k_2}x_{k_2} + c_{k_2}\}. \tag{5}$$

Let $d_{\ell_{k_1},\ell_{k_2}} = \min \|p_i - p_j : p_i \in \ell_{k_1}, p_j \in \ell_{k_2}\|_2$ be the minimum distance between ℓ_{k_1} and ℓ_{k_2} and M_{k_1} the midpoint of ℓ_{k_1}.

1. If $d_{\ell_{k_1},\ell_{k_2}} = 0$ then ℓ_{k_1} is a **linear neighbour** of ℓ_{k_2}, represented in Fig. 5(a)

2. If $d_{\ell_{k_1},\ell_{k_2}} \leq r$ then ℓ_{k_1} is a **radial linear neighbour** of ℓ_{k_2}, represented in Fig. 5(b)
3. if $d_{M_{k_1},\ell_{k_2}} = \min \left\| M_{k_1} - p_j : p_j \in \ell_{k_2} \right\|_2 \leq r$ then ℓ_{k_1} is a **radial midpoint linear neighbour** of ℓ_{k_2}, represented in Fig. 5(c)

In this paper we use three neighbours defined in Definition 1, that is, all three neighbour types are considered neighbours simultaneously.

2.2.2. Linear hotspots

Getis–Ord statistics (Getis and Ord, 2010) are traditionally used to identify statistically significant hotspots by comparing the local sum of a feature and its neighbours with the local sum of all the features. Herein, the features are the rate of events per line segment as in Eq. (3). The adapted Getis–Ord statistic is obtained as

$$G_{ik} = \frac{\sum_{t\in I, s=ik} e_{st} r_{ik}}{\sum_{i=1}^{n} r_{ik}}. \tag{6}$$

and is used to test the hypothesis the null hypothesis of complete spatial randomness (Getis and Ord, 2010) against the alternative hypothesis of a spatial pattern existing, namely clustering. The standardised G_{ik} statistic is given by

$$G^*_{ik} = \frac{\sum_{t\in I, s=ik} e_{st} r_{ik} - \bar{r} \sum_{t\in I, s=ik} e_{st}}{S \sqrt{\frac{n_s \sum_{t\in I, s=ik} (e_{st})^2 - \left(\sum_{t\in I, s=ik} e_{st}\right)^2}{n_s - 1}}} \tag{7}$$

$$\text{where } \bar{r} = \frac{\sum_{k=1}^{k_i} \sum_{i=1}^{n} r_{ik}}{n_s} \tag{8}$$

$$\text{and } S = \sqrt{\frac{\sum_{k=1}^{n_{ik}} \sum_{i=1}^{n} (r_{ik})^2}{n_s} - (\bar{r})^2}. \tag{9}$$

The limiting distribution of G^*_{ik} is a normal distribution.[1,2]

The z-scores and p-values are used to identify *significant* linear hotspots and linear coldspots, as it is the probability of observing such an extreme G^*_{ik} value under the null hypothesis of complete spatial randomness. A feature with a large positive z-score and a small p-value indicates a high degree of spatial clustering, namely a hotspot. A small p-value and a large negative z-score imply a low degree of spatial clustering, namely a coldspot. The higher the clustering, the larger the z-score. At a 95% confidence level a p-value less than 0.05 with a positive z-score implies that the null hypothesis of no spatial pattern is rejected and a significant linear hotspot is present. At a 95% confidence level a p-value less than 0.05 with a negative z-score implies that the null hypothesis of no spatial pattern is rejected and a significant linear coldspot is present.

It is also informative to analyse the spatial distribution of the point pattern over time, that is a spatio-temporal point pattern $\{x_t = \{x_{1,t}, \ldots, x_{2,t}\}\}_{t\in T}$ (Diggle, 2013). The z-scores from the Getis–Ord statistics for each line segment are calculated at chosen time points. The Mann–Kendall test (Kendall and Gibbons, 1990) is applied to the z-scores to analyse the trend in z-scores for each of the line segments and determine temporal hot- and coldspots. Consider the G^*_{ik} statistics for N time points $\{G^*_{ik,t_1}, G^*_{ik,t_2}, \ldots, G^*_{ik,t_N}\}$ for a line segment ℓ_{ik}. The test statistic for the Mann–Kendall test is given by

$$S_{ik} = \sum_{s=1}^{N-1} \sum_{p=i+1}^{N} \text{sign}(G^*_{ik,p} - G^*_{ik,s}), \tag{10}$$

[1] If $G^*_1, G^*_2, \ldots, G^*_n$ is a random sample from a distribution with mean μ and variance $\sigma < \infty$, then the distribution of $Z_n = \frac{\sum_{i=1}^{n} G^*_i - n\mu}{\sigma\sqrt{n}}$ is normal.

[2] The Getis–Ord statistic is a sum of events, therefore by the central limit theorem, if the number of features is large, the limiting distribution of the sum is normal distribution. This therefore allows standardisation to obtain a standardised Getis–Ord function, statistically represented as z-scores. The p-values are then calculated using the z-scores, as well understood in statistics.

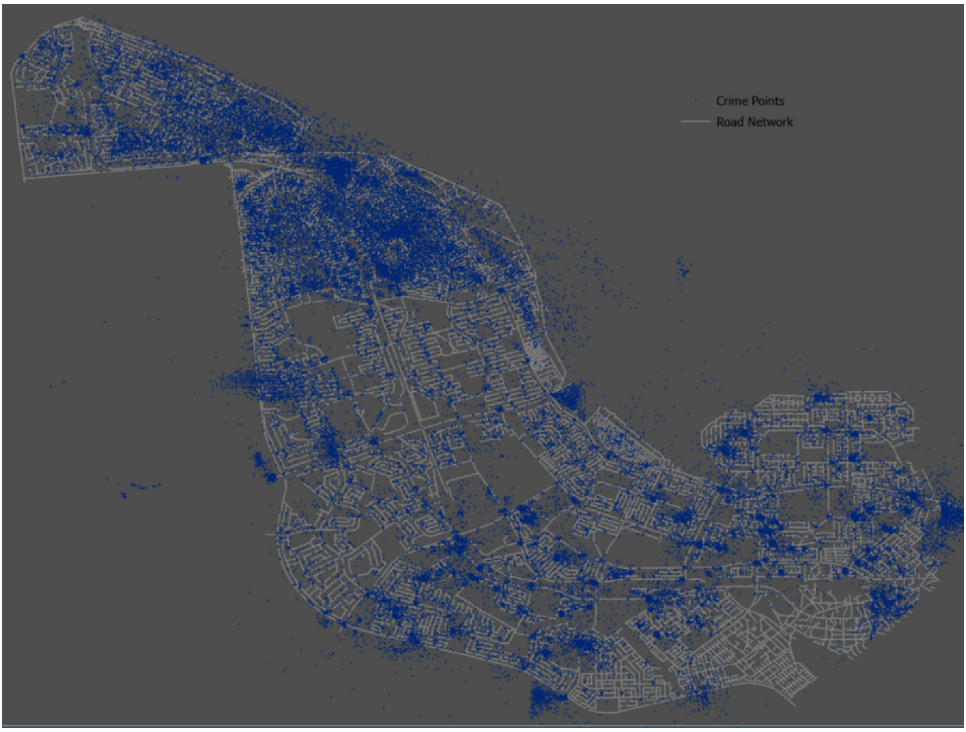


Fig. 6. Crime locations and the road network in Khayelitsha, South Africa 2006–2016. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where

$$\text{sign}(G^*_{ik,p} - G^*_{ik,s}) = \begin{cases} 1 & \text{if } G^*_{ik,s} < G^*_{ik,p} \\ 0 & \text{if } G^*_{ik,s} = G^*_{ik,p} \\ -1 & \text{if } G^*_{ik,s} > G^*_{ik,p}. \end{cases}$$

The z-scores and p-values associated with the test are used to determine if there are statistically significant trends in the data over time. A high negative z-score indicates a downward trend (diminishing hotspots), a high positive z-score indicates an upward trend (emerging hotspots) and a z-score of 0 indicates no trend.

We next apply this new methodology to crime locations in Khayelitsha, South Africa.

3. Results

The data for the study consists of the longitude and latitude geographic locations, and date and time, of crimes reported in Khayelitsha from 2006–2016[3] as well as the road network of Khayelitsha available through ArcGIS[4] For the proposed methodology, the road network represents the linear network and the crime locations represent the point pattern in the vicinity of the road network. Fig. 6 shows the crime locations in blue with the underlying road network in light grey. There are crime locations that fall outside the linear network, these would be excluded in the analysis as they are too far away from the road network.

[3] Ethics approval obtained: NAS208/2019.

[4] ArcGIS [GIS software]. Redlands, CA: Environmental Systems Research Institute, Inc., 2010. www.esri.com.

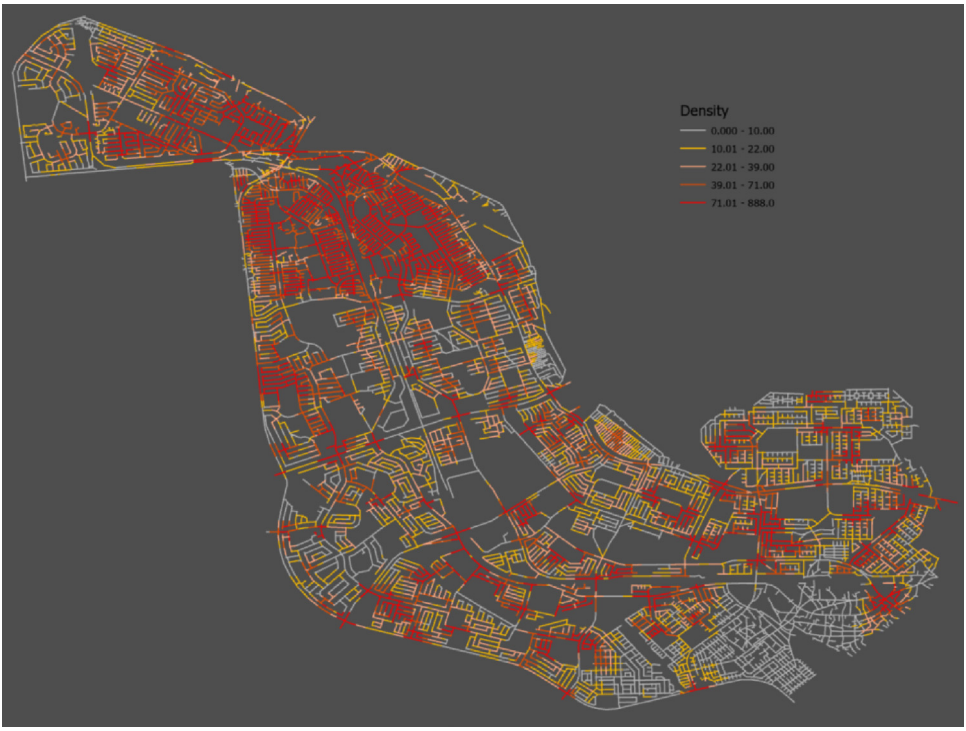


Fig. 7. Visualisation of the rate of events per line segment using Eq. (3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Khayelitsha is a large township located approximately 30 km from the city of Cape Town, South Africa. Khayelitsha was established in 1983 as a late implementation of the Group Areas Act of 1950. The Group Areas Act resulted in the forceful evacuation of Black African families to remote areas distant and distinct from the white urban core and resulted in the complete racial segregation of most of the country's major cities and towns under apartheid. Khayelitsha is one of the poorest[5] and most crime-ridden[6] townships in South Africa with rates of crime consistently above the national average for most types of crime.

This application of our methodology aims to better understand the spatial dimension of crime incident locations in the township and in doing so, provide some insight in informing crime prevention strategies. In particular we aim to investigate if the roads are crime facilitators using the proposed methodology developed herein. Crime facilitators assist criminals in committing crimes or creating disruption (Natarajan et al., 1996). Examples of crime facilitators are cell phones, used to communicate criminal ideas (Natarajan et al., 1996) or guns to overcome resistance during a criminal activity (Griffiths and Chavez, 2004). Streets such as in townships, where there are no street lights, also act as crime facilitators since offenders take advantage of not being easily identified in darker streets (Painter and Farrington, 1997). The methodology in this paper is applied to point patterns in the vicinity of a linear network to investigate street segments as possible crime facilitators. Spatial crime analysis is occurring at increasingly finer levels of aggregation with the street segment now the preferred unit of analysis in assessing crime risk (Braga et al., 2017). Therefore, being able to accurately delineate points (crimes) to line segments (roads) is vital for

5 https://www.weforum.org/agenda/2016/10/these-are-the-worlds-five-biggest-slums/.

6 https://issafrica.org/crimehub.

Fig. 8. Linear hotspot detection using the neighbourhood structures defined in Definition 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

future statistical studies of crime. The concept of crime location in the vicinity of the road network is considered over the concept of crime locations that occur directly *on* a road. To extend the analysis towards points within the vicinity of a linear network, we use the proposed methodology and the new measure of connectivity defined herein, namely neighbours of line segments.

The first step is to determine the spatial distribution of points in the vicinity of a linear network. The adjusted methodology has been applied to the data to obtain the rate of crime per road segment, calculated using the weighted version of Eq. (3). The grid size was determined in a data driven manner based on the average stand size in Khayelitsha. The stands have a uniform structure as the area was originally a forced formal settlement, that is, forced movement of a certain population group to this area and designated housing spaces. A radius of $r = 55$ m was used in determining the line neighbours (the combination of all three line neighbours in Definition 1) considering the denseness of both the road network and the crime point pattern.

Fig. 7 represents the visualisation of the rate of events per line segment (Eq. (3)) of the road network of Khayelitsha, where the road segments are colour coded according to the number of points in their vicinity. The grey line segments represent roads with fewer crimes in their vicinity and the red line segments represent roads with a higher number of crimes in their vicinity.

Next we identify statistically significant hotspots. The proposed method is applied to the data to identify statistically significant hotspots. Fig. 8 represents statistically significant linear hotspots and Fig. 9 represents statistically significant linear coldspots. The red line segments represent roads with a higher than average number of crimes in their vicinity at 90%, 95% and 99% confidence level. The blue line segments represent roads with a lower than average number of crimes in their vicinity 90%, 95% and 99% confidence level.

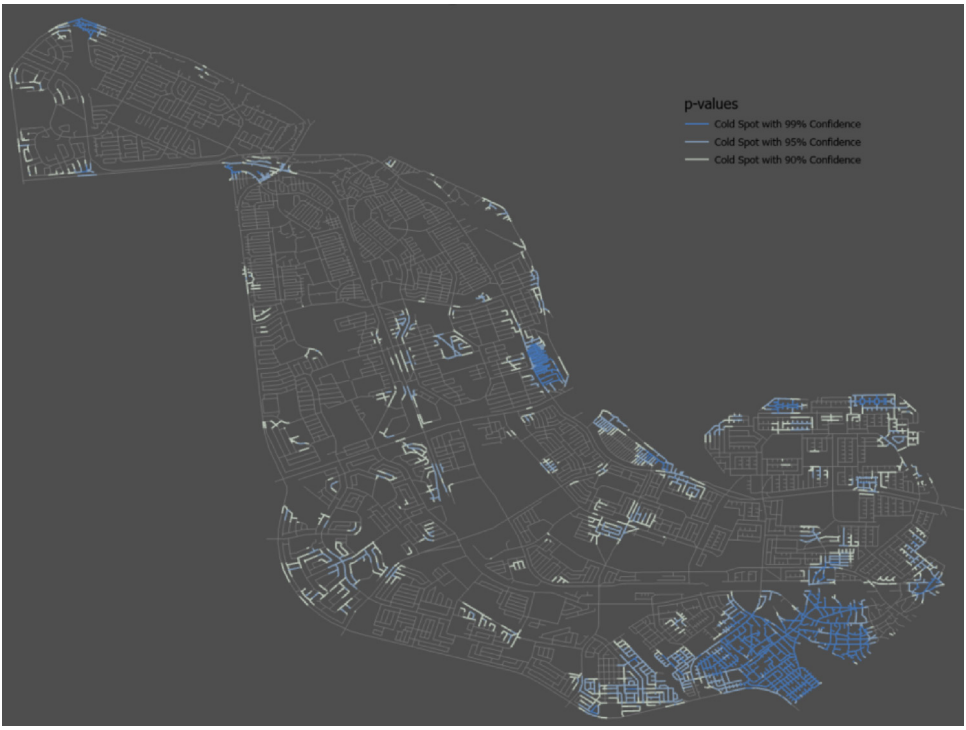
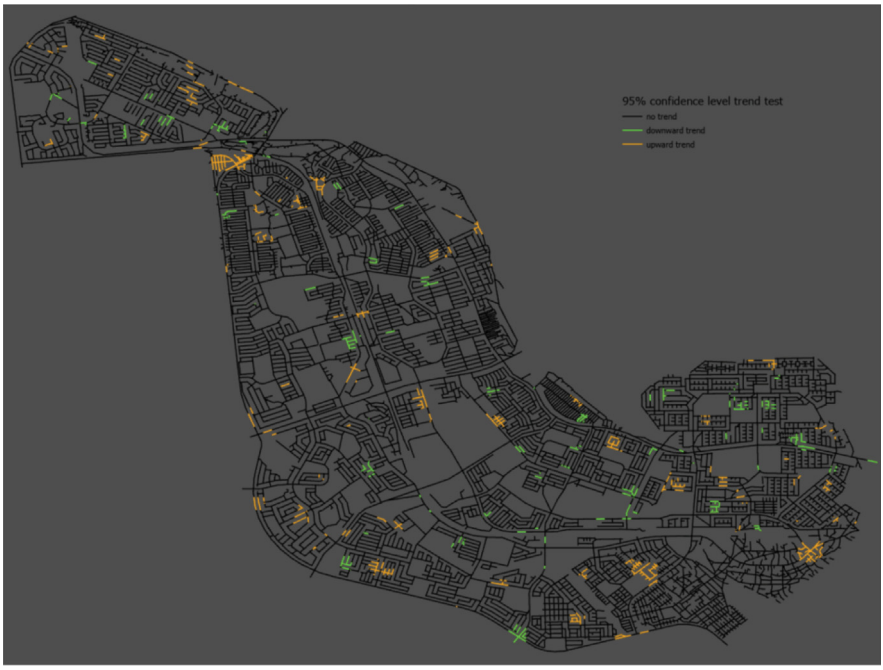


Fig. 9. Linear coldspot detection using the neighbourhood structures defined in Definition 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

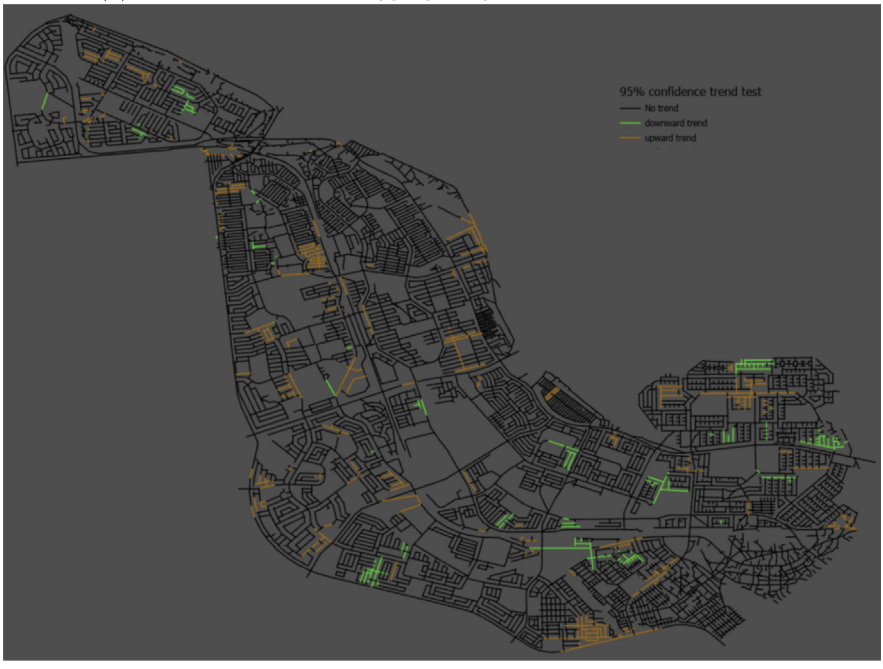
So far results take into account the spatial aspect of the data but ignore the temporal aspect. It is informative to analyse the spatial–temporal distribution of the data. A trend test is conducted to analyse the 6 monthly and yearly trends, which is possible since the exact times for the crime occurrences are known. The Mann–Kendall test is applied to the Getis–Ord statistics obtained for each line segment over the time period. In Fig. 10, the brown line segments represent line segments with an upward trend and green lines represent line segment with a downward trend for a 6 month time interval (Fig. 10(a)) and for a yearly interval (Fig. 10(b)).

We further investigate the spatial relationship between the detected linear hot- and coldspots by making use of the multi-K (also known as the cross-K function) (Lotwick and Silverman, 1982). In Fig. 11, the blue curve represents the multi-K function calculated using the data in our study, while the black curve represents a simulated multi-K function in the case where the two point pattern cluster around each other with the confidence interval represented by the red curves. If the blue line falls between the confidence interval, then it implies that the coldspots cluster around the hotspots. Fig. 11 shows that the two types in fact inhibit each other. The blue line represents the multi-K function calculated using the data in our study, while the black line represents a simulated multi-K function in the case where the two point pattern cluster around each other with the confidence interval represented by the red lines. If the blue line were to lie in the confidence interval, it would imply that the coldspots cluster around the hotspots.

We next provide a validation of this new methodology. Fig. 12(a) shows traditional hotspot detection ignoring the linear network. This uses the same grid as the grid used to determine the line segments ℓ_{ik} to allow for a fair comparison. To compare the new methodology to the traditional approach, we display the differences in the p-values, using the average per grid cell for the linear hotspot method since there are a number of line segments per cell. We also display the variances



(a) Crime trends when aggregating at 6 month time intervals



(b) Crime trends when aggregating at yearly time intervals

Fig. 10. Emerging and diminishing linear hot- and cold-spots over aggregated time intervals of 6 months and 1 year. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

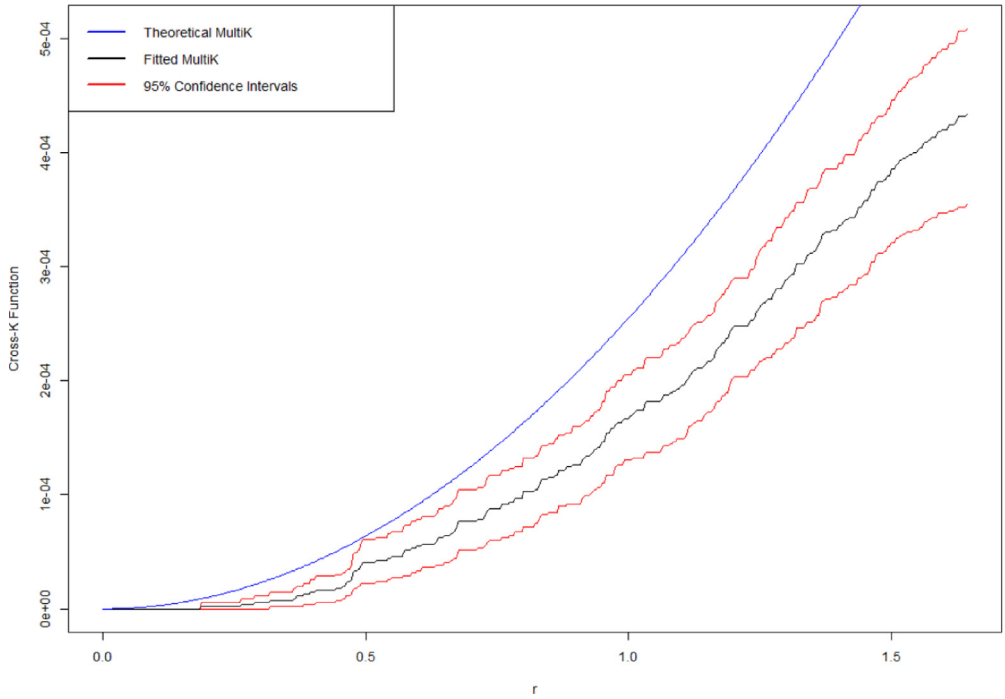


Fig. 11. Multi-K function between the linear hot- and coldspots showing evidence of inhibition. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

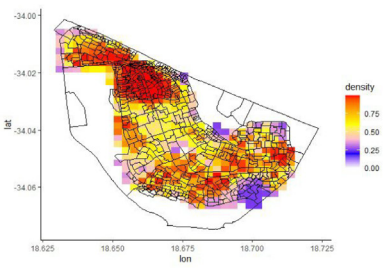
of these p-values per grid cell in Fig. 12(c), the maximum p-values per cell in Fig. 12(d) and the minimum p-values per cell in Fig. 12(e).

In Fig. 12(b), the far right indicates some large p-value differences. Fig. 13 provides a closer look at this. Fig. 13(a) shows the traditional hotspots with the crime locations. There are some crimes which are not in a grid cell due to no formal road network present there, resulting in the discrepancies between the two approaches. Fig. 13(b) shows the linear hotspot average p-values. Fig. 13(c)–(e) show the variances, maximum and minimum values as well. In Fig. 13(f) the linear hotspots are displayed.

4. Discussion

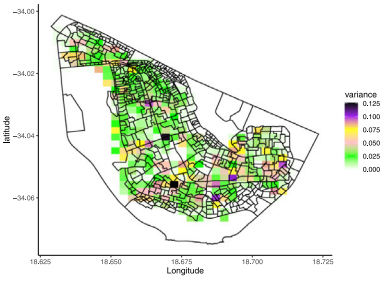
The proposed methodology presents a new definition of hotspots on a linear network, termed linear hotspots. This includes a new definition of neighbours for line segments. We present an application of this new methodology on crime rates in a large township in South Africa. The study shows, in Fig. 8, line segments (in red) with the higher than average number of crimes in their vicinity as compared to neighbouring line segments. Fig. 9 shows line segments (in blue) with the lower than average number of crimes in their vicinity as compared to neighbouring line segments. The Getis–Ord statistics and the corresponding z-scores for each of the line segments are calculated and the statistically significant linear hotspots are obtained at a 95% confidence level. These line segments are thus the linear crime hotspots and coldspots. In our study we noticed that most of these occur on or near areas where roads intersect and that they are then clustered together, namely that linear hotspots are spread over the road network but occur locally. This crime phenomenon has also recently been observed in Andresen and Malleson (2013).

In the available data the crime locations have been recorded according to their date and time. Herein we also present methodology for how linear hotspots change over time. The implementation



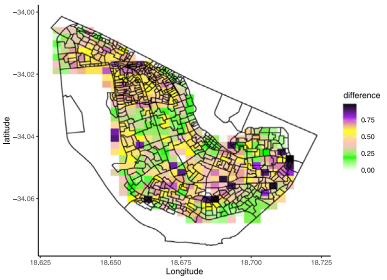
(a) Traditional hotspot detection per grid cell using the Getis-Ord statistic

(b) Differences per grid cell between the p-values of (a) and linear hotspot average p-values



(c) Variances of the differences in (b) per grid cell

(d) Maximum difference per grid cell between the p-values of (a) and the linear hotspot p-values

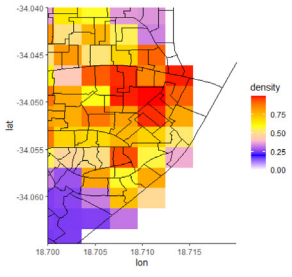


(e) Minimum difference per grid cell between the p-values of (a) and the linear hotspot p-values

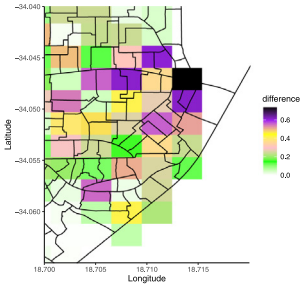
Fig. 12. Validation of the proposed methodology using the average, variance, minimum and maximum of the difference in p-values compared to the p-values from traditional hotspot analysis on a grid.

in Fig. 10 visualises areas with emerging (red) and diminishing hotspots (blue). Using finer time intervals e.g. 6 months, picks up fewer hot- and coldspots. This could indicate a good sensitivity to the data. The study could be extended by investigating linear hotspots of each crime type in order to identify the type of crimes that cluster in specific areas. It is likely that different crime types exhibit different spatio-temporal characteristics with respect to a linear network. These results are useful for law enforcement for tracking of hotspots and to investigate if resources allocated to the areas are useful i.e. to investigate if there is an increase or decrease in crime in areas of interest over the specified time period, and if a change in resource allocation is needed.

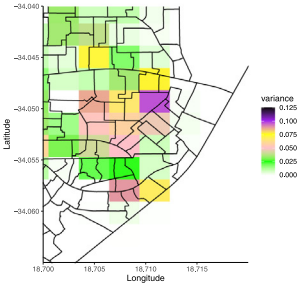
With the linear hotspot methodology development in place, the results were validated. Traditionally, prior information about an area is used to identify the validity of the results. In this case, no prior information is available and the results can only be compared to traditional hotspot detection. The traditional hotspot technique identifies hotspots at the macro-level, namely grid cells, while the linear hotspot technique identifies linear hotspots at the micro-level, namely roads within a



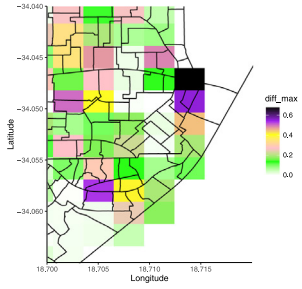
(a) Traditional hotspot detection per grid cell using the Getis-Ord statistic



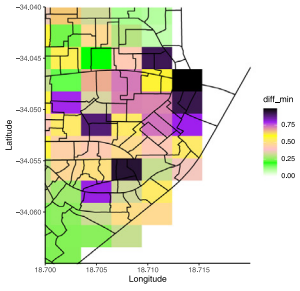
(b) Differences per grid cell between the p-values of (a) and linear hotspot average p-values



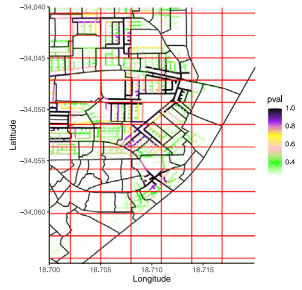
(c) Variances of the differences in (b) per grid cell



(d) Maximum difference per grid cell between the p-values of (a) and the linear hotspot p-values



(e) Minimum difference per grid cell between the p-values of (a) and the linear hotspot p-values



(f) The linear hot- and coldspots detected

Fig. 13. A closer look at the South East corner of Khayelitsha 12.

grid cell. This is visible in Fig. 12. The proposed methodology provides a finer level of detection by taking into account the road network, which allows for micro-level determination of what generated the hotspot. Note that the consistent coldspots detected around the boundary of Khayelitsha in the traditional hotspot technique are emphasised due to edge effects. Nonetheless, these are still informative as there is little crime there as can be seen in Fig. 6. The validation results in Figs. 12 and 13 indicates that the proposed methodology captures the same information as traditional hotspots, visible by the low variances and differences, but captures finer scale information visible at line segment level.

The proposed methodology can also be applied to cases of curved roads and curved road segments, without any significant changes besides checking that distances are calculated correctly. We did not consider the impact of in-homogeneous distribution of line segments across the spatial domain. The techniques in Steenberghen et al. (2010, 2004) should be considered for extensions to take this into account.

5. Conclusion

This study provides for the first time a method for detecting linear hotspots for points in the vicinity of a linear network that do not necessarily fall on the linear network. This method is motivated by crime locations amongst a road network. It effectively uses a linear network as additional spatial structural information to a spatial point pattern. The detected inhibition between the linear hot- and coldspots opens up further discussion on covariates. It is likely the location of a linear hot- or coldspot also relates to the covariates in its vicinity, rather than to the location of another linear hot- or cold spot nearby. Future work will investigate inclusion of covariates, such as population density, into the methodology.

In this paper we identified road networks as crime facilitators. We used maps showing differences of p-values from linear network methods and traditional methods to validate our results. The hypothesis is that most linear hotspots are contained in the traditional hotspots since the road networks serve as crime facilitators. We explored the grid size used for splitting the linear network in line segments in calculating traditional hotspots such that the two methods are calculated at the same level of scale.

Identification and characterisation of linear hotspots are important tools for law enforcement agencies in order to better understand the pattern and genesis of crimes, and thus to reduce the crime rates and improve livability in an urban area. This is phrased as one of the sustainable development goals, namely SDG 16,[7] which aims to significantly reduce all forms of violence, and work with governments and communities to find lasting solutions to conflict and insecurity. The method presented in our study integrates graph theory and spatial statistics for detecting linear hotspots among points that occur in the vicinity of the network, incorporating the linear network as a covariate for the spatial point pattern occurrence. For the specific application in Khayelitsha, the pattern showed several linear hotspots that could be well interpreted and potentially help government organisations to allocate measures that reduce criminality. Spatial hotspot determination was originally motivated exactly for crime analysis (Chakravorty, 1995). Further research should test the sensitivity of the methodology to the optimal grid size and the radius used in determining linear neighbours. The proposed methodology is also suited for wider use where the aim is to detect patterns in points in the vicinity of a network.

A road network is a common linear network, but there are other types of linear networks, for example, river networks or flight paths. The methodology developed in this paper is also not only limited to crime points in the vicinity of a road network. It can be applied to cases such as the location of trees on or along the river network (Spooner et al., 2004), the use of social network analysis in veterinary epidemiology to model the spread of animal diseases in different farms (Dubé et al., 2011) and determining how animals move within their environment by analysing the interconnectivity of the locations they move between (Jacoby et al., 2012). The general methodology thus has many alternative applications as well.

The proposed methodology only considered how point patterns in a linear network change over time when the linear network stays constant. The methodology can be extended to analyse spatio-temporal distribution of point pattern in the vicinity of a linear network and how the change in the linear network structure over time affects the spatial distribution of the point pattern.

Acknowledgements

This work is based upon research supported by the National Research Foundation, South Africa (Research chair: Computational and Methodological Statistics, Grant number 71199). Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF. Acknowledge is also given to ESRI South Africa for financial support.

[7] https://sustainabledevelopment.un.org/?menu=1300.

References

Andresen, M., Malleson, N., 2013. Crime seasonality and its variations across space. Appl. Geogr. 43, 25–35.

Ang, Q., Baddeley, A., Nair, G., 2012. Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. Scand. J. Stat. 39 (4), 591–617.

Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27 (2), 93–115.

Anselin, L., 2019. A local indicator of multivariate spatial association: extending Geary's C. Geogr. Anal. 51 (2), 133–150.

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., Davies, T.M., 2021. Analysing point patterns on networks—A review. Spatial Stat. 42, 100435.

Borruso, G., 2005. Network density estimation: analysis of point patterns over a network. In: International Conference on Computational Science and Its Applications. Springer, pp. 126–132.

Borruso, G., 2008. Network density estimation: a GIS approach for analysing point patterns in a network space. Trans. GIS 12 (3), 377–402.

Braga, A., Andresen, M., Lawton, B., 2017. The law of crime concentration at places. J. Quant. Criminol. 33, 421–426.

Chakravorty, S., 1995. Identifying crime clusters: The spatial principles. Middle States Geogr. 28, 53–58.

Comas, C., Costafreda-Aumedes, S., Lopez, N., Vega-Garcia, C., 2019. On the correlation structure between point patterns and linear networks. Spatial Stat. 29, 192–203.

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

D'Angelo, N., Adelfio, G., Mateu, J., 2021. Assessing local differences between the spatio-temporal second-order structure of two point patterns occurring on the same linear network. Spatial Stat. 45.

Diggle, P., 2013. Statistical Analysis of Spatial and Spatio-TEmporal Point Patterns. Chapman and Hall/CRC.

Dubé, C., Ribble, C., Kelton, D., McNab, B., et al., 2011. Introduction to network analysis and its implications for animal disease modelling. Revue Sci. Et Tech.-OIE 30 (2), 425.

Eckardt, M., Mateu, J., 2016. Structured network regression for spatial point patterns. arXiv:1607.06685.

Eckardt, M., Mateu, J., 2018. Point patterns occurring on complex structures in space and space-time: An alternative network approach. J. Comput. Graph. Statist. 27 (2), 312–322.

Ejigu, B.A., Wencheko, E., 2020. Introducing covariate dependent weighting matrices in fitting autoregressive models and measuring spatio-environmental autocorrelation. Spatial Stat. 38, 100454.

Getis, A., Ord, J., 1995. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24 (3).

Getis, A., Ord, J., 2010. The analysis of spatial association by use of distance statistics. In: Perspectives on Spatial Data Analysis. Springer, pp. 127–145.

Griffiths, E., Chavez, J.M., 2004. Communities, street guns and homicide trajectories in Chicago, 1980–1995: Merging methods for examining homicide trends across space and time. Criminology 42 (4), 941–978.

Jacoby, D.M., Brooks, E.J., Croft, D.P., Sims, D.W., 2012. Developing a deeper understanding of animal movements and spatial dynamics through novel application of network analyses. Methods Ecol. Evol. 3 (3), 574–583.

Jiang, B., Okabe, A., 2014. Different ways of thinking about street networks and spatial analysis. Geogr. Anal. 46 (4), 341–344.

Kendall, M., Gibbons, J.D., 1990. Rank Correlation Methods, fifth ed. A Charles Griffin Title.

Lotwick, H.W., Silverman, B.W., 1982. Methods for analysing spatial processes of several types of points. J. R. Stat. Soc. Ser. B Stat. Methodol. 44, 406–413.

Mateu, J., Moradi, M., Cronie, O., 2020. Spatio-temporal point patterns on linear networks: Pseudo-separable intensity estimation. Spatial Stat. 37, 100400.

Moradi, M.M., Rodríguez-Cortés, F.J., Mateu, J., 2018. On kernel-based intensity estimation of spatial point patterns on linear networks. J. Comput. Graph. Statist. 27 (2), 302–311.

Natarajan, M., Clarke, R.V., Belanger, M., 1996. Drug dealing and pay phones: The scope for intervention. Secur. J. 7 (4), 245–251.

Okabe, A., Sugihara, K., 2012. Spatial Analysis Along Networks: Statistical and Computational Methods. John Wiley & Sons.

Painter, K., Farrington, D.P., 1997. The crime reducing effect of improved street lighting: The dudley project. Situat. Crime Prev.: Success. Case Stud. 2, 209–226.

Peeters, A., Zude, M., Käthner, J., Ünlü, M., Kanber, R., Hetzroni, A., Gebbers, R., Ben-Gal, A., 2015. Getis-Ord's hot-and cold-spot statistics as a basis for multivariate spatial clustering of orchard tree data. Comput. Electron. Agric. 111, 140–150.

Rakshit, S., Nair, G., Baddeley, A., 2017. Second-order analysis of point patterns on a network using any distance metric. Spatial Stat. 22, 129–154.

Ramos, R.G., Silva, B.F., Clarke, K.C., Prates, M., 2021. Too fine to be good? Issues of granularity, uniformity and error in spatial crime analysis. J. Quant. Criminol. 37 (2), 419–443.

Rasmussen, J.G., Christensen, H.S., 2021. Point processes on directed linear networks. Methodol. Comput. Appl. Probab. 23 (2), 647–667.

Shino, S., 2008. Analysis of a distribution of point events using the network-based quadrat method. Geogr. Anal. 40 (4), 380–400.

Songchitruksa, P., Zeng, X., 2010. Getis-Ord Spatial Stat. to identify hot spots by using incident management data. Transp. Res. Rec. 2165 (1), 42–51.

Spooner, P., Lunt, I., Okabe, A., Shiode, S., 2004. Spatial analysis of roadside acacia populations on a road network using the network K-function. Landsc. Ecol. 19 (5), 491–499.

Steenberghen, T., Aerts, K., Thomas, I., 2010. Spatial clustering of events on a network. J. Transp. Geogr. 18 (3), 411–418.

Steenberghen, T., Dufays, T., Thomas, I., Flahaut, B., 2004. Intra-urban location and clustering of road accidents using GIS: a Belgian example. Int. J. Geogr. Inf. Sci. 18 (2), 169–181.

Suryowati, K., Bekti, R., Faradila, A., 2018. A comparison of weights matrices on computation of dengue spatial autocorrelation. In: IOP Conference Series: Materials Science and Engineering, Vol. 335. (1), IOP Publishing, 012052.

Tompson, L., Partridge, H., Shepherd, N., 2009. Hot routes: Developing a new technique for the spatial analysis of crime. Crime Mapp.: A J. Res. Pract. 1 (1), 77–96.

Yamada, I., Thill, J., 2003. Empirical comparisons of planar and network K-functions in Traffic Accident Analysis. In: Proceedings of the 82nd Transportation Research Board Annual Meeting, pp. 2–5, Washington DC.

Yamada, I., Thill, J., 2007. Local indicators of network-constrained clusters in spatial point patterns. Geogr. Anal. 39 (3), 268–292.